CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Comparative microbial genome analysis of lactobacilli
Authors	Harris, Hugh
Publication date	2017
Original Citation	Harris, H. 2017. Comparative microbial genome analysis of lactobacilli. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2017, Hugh Harris. - http://creativecommons.org/licenses/by-nc-nd/3.0/
Download date	2026-11-15 01:09:58
Item downloaded from	https://hdl.handle.net/10468/6086

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh



University College Cork

Coláiste na hOllscoile Corcaigh

Comparative microbial genome analysis of lactobacilli

A Thesis Presented to the

National University of Ireland

for the

Degree of Doctor of Philosophy

By

Hugh Harris, M.Sc.

School of Microbiology

National University of Ireland

Cork

Supervisors:

Prof. Paul O'Toole

Dr Marcus Claesson

Head of School:

Prof. Gerald Fitzgerald

November 2017

Dedicated to young Hugh

Things get better

“Long is the way and hard, that out of Hell leads up to Light.”

- John Milton (Paradise Lost)

Bart Simpson: Um, Dad?
Homer Simpson: Yeah?
Bart Simpson: What is the mind? Is it just a... system of impulses**,** or... is it... something tangible?
Homer Simpson: Relax! What is mind? No matter. What is matter? Never mind.
Bart Simpson: Thanks, Dad.
Homer Simpson: Good night, son.

- The Simpsons

***

“First they take the dingle-bop and they smooth it out with a bunch of schleem. The schleem is then repurposed for later batches. They take the dingle-bop and they push it through the grumbo where the fleeb is rubbed against it. It's important that the fleeb is rubbed because the fleeb has all of the fleeb juice. Then, a schlami shows up and he rubs it and spits on it. They cut the fleeb. There's several hizzards in the way. The blamphs rub against the chumbles, and the plubis and grumbo are shaved away. That leaves you with a regular old plumbus.”

- Rick and Morty (How plumbuses are made)

***

Clark: “Yeah, but I will have a degree. You’ll be serving my kids fries at a drive through on our way to a skiing trip.”
Will Hunting: “That may be. But at least I won’t be unoriginal.”

- Good Will Hunting (Bar scene)

TABLE OF CONTENTS

Declaration 5
Chapter I 6
Chapter II 75
Chapter III 113
Chapter IV 147
Chapter V 181
Acknowledgements 192
Appendix 194

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Signed: _____

Hugh Harris

November 2017

Chapter I

Literature Review

TABLE OF CONTENTS

1 PREFACE 9

1.1 GENOMIC DATA 9

1.2 SCOPE OF THE REVIEW 10

2 MICROBIAL GENOMICS 11

2.1 A HISTORICAL PERSPECTIVE 11

2.1.1 MICROBIOLOGY 11

2.1.2 EVOLUTION 12

2.1.3 GENETICS 13

2.1.4 MATHEMATICS 14

2.1.5 COMPUTER SCIENCE 15

2.2 GENOME SEQUENCING 16

2.2.1 FIRST-GENERATION SEQUENCING 17

2.2.2 NEXT-GENERATION SEQUENCING 18

2.2.3 THIRD-GENERATION SEQUENCING 21

2.3 GENOME ASSEMBLY 22

2.3.1 PRE-ASSEMBLY QUALITY CONTROL 23

2.3.2 ASSEMBLY ALGORITHMS 24

2.3.3 ASSEMBLY STATISTICS 26

2.4 GENOME ANNOTATION 27

2.4.1 GENE PREDICTION 28

2.4.2 FUNCTIONAL ANNOTATION 32

3 COMPARATIVE GENOMIC STUDIES AND EMERGING CONCEPTS 36

3.1 EARLY COMPARATIVE GENOMIC STUDIES IN LACTOBACILLI 36

3.2 THE PAN-GENOME 37

3.2.1 HOMOLOGY, ORTHOLOGY AND PARALOGY 37

3.2.2 THE EUKARYOTIC PAN-GENOME 39

3.2.3 THE PROKARYOTIC PAN-GENOME 40

3.2.4 THE PAN-GENOMES OF SPECIES WITHIN THE *L. CASEI* GROUP 41

3.2.5 THE PAN-GENOMES OF OTHER *LACTOBACILLUS* SPECIES 43

3.3 HORIZONTAL GENE TRANSFER AND GENE LOSS 45

3.3.1 RECOMBINATION 45

3.3.2 PLASMIDS 46

3.3.3 BACTERIOPHAGES 48

3.3.4 TRANSPOSASES 51

3.3.5 GENE LOSS 53
3.3.6 STRAIN-SPECIFIC GENES 55
3.4 THE PARAPHYLETIC NATURE OF *LACTOBACILLUS* 56
3.4.1 MONOPHYLY, POLYPYLY AND PARAPHYLY 56
3.4.2 *LACTOBACILLUS* PHYLOGENY 57
3.5 EVOLUTIONARY RATE 60
3.5.1 MUTATION RATE 60
3.5.2 SYNONYMOUS AND NON-SYNONYMOUS MUTATIONS 62
3.5.3 SELECTION PRESSURE AND EVOLUTIONARY RATE 63
4 BIBLIOGRAPHY 66

1 PREFACE

1.1 GENOMIC DATA

The bacteriophage MS2, an RNA-based virus that infects *E. coli*, was the first organism to have its genome made digitally available using sequencing technology in 1976. Following this milestone, the sequencing of more complex genomes from PhiX174 to *Drosophila melanogaster* culminated in the sequencing of the human genome, which was completed in 2003. This was a thirteen-year project that was undertaken by Celera Genomics (a private company led by Craig Venter) and an NIH-funded public initiative led by Francis Collins (Moraes and Goes, 2016).

Shortly after the human genome had been sequenced, Craig Venter and his team went on an expedition to the Sargasso Sea in a sailing boat. They sampled the microbial content of the area at multiple sites by passing water through filters and by sequencing the resulting residue. They discovered a surprising amount of microbial diversity estimated at 1,800 species containing 1.2 million previously unknown genes (Venter et al., 2004).

Sequencing projects like the Sargasso Sea expedition reveal an impressive level of diversity that was considerably underestimated using purely culture-based methods. Improvements in technology and an increase in the sequencing and analysis of molecular data have propelled us into an era of biological research where important insights are no longer dominated by culture-dependent methods. The field of Bioinformatics is rapidly expanding, created by the fusion of Biology, Computer Science and Mathematics. This multi-disciplinary approach to research brings with it a new kind of research, applying computer algorithms and statistical knowledge to a wide variety of biological questions. The size of some sequence datasets is so immense that a team of researchers could not manually read through even a small fraction of the data over the course of their lives. Yet sub-disciplines of Bioinformatics such as Comparative Genomics have the power to probe a seeming ocean of nucleotide bases for patterns that expand on current knowledge, strengthening theoretical insights with high-throughput empirical data.

1.2 SCOPE OF THE REVIEW

This literature review focuses on comparative microbial genome analysis in lactobacilli. *Lactobacillus* species are involved in food fermentation, probiotics and starter cultures for dairy products, although they can have negative roles too like in cavity formation due to acid production in the presence of sugar (Salvetti et al., 2012). The paraphyletic nature of lactobacilli and the historical misclassification of species due to contradictory genotype/phenotype sub-groupings make this large bacterial division an interesting and a challenging task for present and future bioinformaticians.

The sub-discipline of Comparative Genomics relies on the comparison of functional and phylogenomic properties of multiple genomes. Studies have been conducted on as few as two genomes of the same strain to hundreds of genomes scattered across the tree of life. This review will bring together relevant literature involving the comparison of multiple genomes of *Lactobacillus* and will provide a comprehensive description of the accumulation of knowledge since the first studies to the present day.

The structure of this review will proceed through the origin and history of Microbial Genomics to a literature review of comparative microbial genome analysis in lactobacilli, focusing on key concepts and insights. Early sections give more general overviews, but microbes are focussed on where possible. Relevant software and bioinformatic techniques are included where they add to the understanding of the topics being covered.

2 MICROBIAL GENOMICS

2.1 A HISTORICAL PERSPECTIVE

Microbial Genomics is possible only because it is built on a history of older and more established disciplines. It is easy to take the following facts for granted: bacteria exist; they vary functionally relative to other, closely related bacteria; these closely related bacteria all evolved from a common ancestor; functional variation is due to digital, heritable variations in nucleotide sequences; these variations can be analysed using machines capable of sophisticated, high-speed computations. These facts arise from five separate disciplines - Microbiology, Evolution, Genetics, Mathematics and Computer Science - which have been combined together into a multidisciplinary approach to modern biological research.

2.1.1 MICROBIOLOGY

The microscopic fruiting bodies of mould were first observed in 1665 by Robert Hooke and, in 1676, Antonie van Leeuwenhoek observed the first bacteria using a single-lens microscope of his own design. The existence of organisms too small to be seen with the naked eye had been hypothesised since ancient times, but the drawings and descriptions of Hooke and Leeuwenhoek foreshadowed the emergence of the field of Microbiology (Lane, 2015). The fact that microbes exist at all opens up a vast array of questions into exactly what it is they do and what effects they have on human health and wellbeing.

Microbiology became firmly established as a science in the 1800s with the pioneering work of Ferdinand Cohn, Robert Koch and Louis Pasteur. Cohn developed the first taxonomic classification of bacteria based on their shape - a scheme that is still in use today. Pasteur disproved the theory of spontaneous generation with a series of ingenious experiments showing that meat broth remains sterile when air-borne bacteria are prevented from reaching it. Robert Koch solidified the germ theory of disease by applying a set of rules to observations while

isolating and re-introducing disease-causing bacteria to susceptible hosts, a series of logical steps that became known as Koch's postulates (Madigan, 2012).

No longer could microbes spontaneously arise from nothing; they were pre-existing lifeforms that were amenable to categorisation and a specific microbe had unique properties that caused a specific disease. The subsequent discoveries of Martinus Beijerenk (discoverer of viruses and enrichment culture techniques), Sergei Winogradsky (discoverer of the role of bacteria in geochemical processes) and many others led to a rapid expansion of the field of Microbiology and therefore a considerable increase in the number of implications that further research might reveal (Dworkin, 2012).

2.1.2 EVOLUTION

As the field of Microbiology blossomed, Charles Darwin was writing his book *On the Origin of Species*. Darwin's book, published in 1859, expounded the theory of evolution by means of natural selection. The idea that a species could change over time had existed since ancient Greece, and many hypotheses were put forward over the centuries to try to explain the forces behind evolution. Most notably, Jean-Baptiste Lamarck proposed that organisms arose through spontaneous generation, became increasingly complex with the passing of generations and adapted to their environment through the inheritance of acquired characteristics (Burkhardt, 2013).

Just as Pasteur dispensed with the belief that the spontaneous generation of life is commonplace, Darwin showed that adaptive variations are selected by the environment and less "fit" individuals fail to survive and reproduce. Importantly, he provided convincing arguments (each one strengthened by detailed observations spanning decades) that a species can give rise to new species and he speculated that all life arose from a single common ancestor in the distant past.

The hierarchical nature of species was propounded by Carl Linnaeus in 1735, and the categorisation of microbes began with Cohn in the 1800s, but the idea that all species are essentially cousins of each other with varying evolutionary distances was revolutionary. Applied to Microbiology, Darwin's theory portrayed microbes as organisms that change and adapt to their environments.

Natural selection transforms Biology from a collection of processes and facts into a cohesive scientific field held together by a fundamental theory that explains the complexity of life. A famous essay by Theodosius Dobzhansky summarises the impact of Darwin's work: *Nothing in Biology Makes Sense Except in the Light of Evolution* (Dobzhansky, 1973).

2.1.3 GENETICS

While Darwin's theory explained adaptive change in Biology, it lacked a mechanism for heredity – how are adaptive traits passed down through the generations? Darwin didn't rule out Lamarckian inheritance, which postulated a blending of characteristics from both parents. It was Gregor Mendel, an Augustinian friar and contemporary of Darwin, conducting experiments on pea plants in his garden in Brno, who developed the concept of units of inheritance and fathered the science of Genetics (Mendel, 1866)

Mendel focussed on seven traits of pea plants, meticulously recording the number of each type (e.g. green versus yellow seeds) in the offspring of crossed parents. His experiments gave rise to the idea of dominant and recessive traits as well as the independent assortment of phenotypes in successive generations of plants (Slack, 2014). The supposed blending mechanism of inheritance proposed by Lamarck and others and partly supported by Darwin was largely ruled out by Mendel's mathematical treatment of heritable traits – his results suggested that traits were particulate and inherited as individual units.

Mendel's work fell into obscurity but was rediscovered late in the 1800s by scientists working on related phenomena. The subsequent decades saw the dawn of molecular genetics as the search for the molecules responsible for inheritance escalated. The position of genes on chromosomes was suggested by Thomas Hunt Morgan based on observational data from mutations in fruit flies and the linear arrangement of genes on chromosomes was demonstrated by his student, Alfred Sturtevant, in 1913 (Sturtevant, 1913). The molecular nature of the gene was solidified with the Avery-MacLeod-McCarty experiment in 1944, which demonstrated that DNA was the molecule that carried genetic information in bacterial transformation (Avery et al., 1944). James Watson and Francis Crick

determined the double-helical structure of DNA in 1953 using the x-ray crystallography results of Rosalind Franklin, elucidating its reverse complementarity and showing that adenine always binds with thymine and guanine with cytosine (the infamous A, T, G and C nucleotide bases of genetics) (Watson and Crick, 1953). The discovery of DNA led to a flurry of further research into the genetic code, the set of rules that determine the translation of nucleotides into proteins. A microorganism, just like a fruit fly or a human being, could now be thought of as the phenotypic representation of digital, discrete genetic instructions that are subject to evolution by natural selection. The integration of multiple scientific disciplines was leading to a more detailed understanding of the complex nature of life. This process was to continue.

2.1.4 MATHEMATICS

Mathematics and Biology are traditionally two entirely separate disciplines and it was originally thought that Mathematics would not make much of a contribution to biological research due to the messy complexity of life compared to the deterministic predictability of mathematical equations and formulas. However, the integration of evolutionary theory with Genetics led by Ronald Fisher, used a mathematical framework to create what is now known as the modern synthesis, a combination of the ideas of Darwin and Mendel. In his book of 1930, *The Genetical Theory of Natural Selection*, Fisher showed that the appearance of continuous variation can be explained by the interaction of multiple discrete genetic units, a controversial idea at the time. This mathematical emphasis on evolutionary ideas used by Fisher was implemented by J.B.S Haldane as he quantified natural selection in peppered moths in the case of industrial melanism (where a dark colouration is selectively favoured due to soot deposits on trees). Sewell Wright, too, followed suit by studying complexes of interacting genes and he introduced the concept of an adaptive landscape in 1932 where adaptive peaks of different heights could be bridged in small populations through genetic drift. Fisher, Haldane and Wright together created the field of population genetics, a discipline infused with concepts from evolutionary and ecological theory, the principles of Genetics and substantial mathematical theory (Slack, 2014). The application of sub-disciplines of

Mathematics, particularly Probability and Statistics, directly to the nucleotide and amino acid composition of biomolecules would require the integration of another discipline.

2.1.5 COMPUTER SCIENCE

Devices such as the abacus have been used to aid computation for thousands of years, but the origin of Computer Science rests on several theoretical discoveries and their incorporation into the building of machines capable of computation. Charles Babbage developed the concept of the first programmable computer, the difference engine, in the early 1800s as an aid to navigation. In 1833, he expanded the concept into the analytic engine, a general-purpose mechanical computing device that took punched cards as input. The analytic engine was never completed, but a simplified working version was built years after his death (Haas, 1994). George Boole, an English mathematician, invented binary algebra in the mid-1800s, which is the conceptual foundation of logic gates that form the building blocks of all modern computers (Boole, 1847). Alan Turing in his paper of 1936, *On Computable Numbers*, introduced the concept of the stored program, where all computational instructions are stored in memory (Turing, 1936). Before this, a computer program was fixed in hardware and the introduction of a new program involved the re-wiring of the machine. Turing's "universal computing machines" led to the computational flexibility of today's computers.

The digital computer evolved from using vacuum tubes, to transistors to integrated circuits, becoming faster and smaller following Moore's Law, which states that the processing power of computers will double every two years (Dasgupta, 2016). The digital representation of molecular sequences and the power of modern computers have created a new laboratory for research, one that replaces bench science equipment and an array of chemicals with biological data and computer logic. Leeuwenhoek's microscopes allowed him to discover a new world beyond the limits of what we can see; so too will the computer expand our understanding of nature, allowing us to discern patterns and processes that were previously embedded in biological phenomena too vast and interconnected to be understood. The computer has truly become an essential tool for scientists.

Microbial Genomics is a hybrid of disciplines that stretch back over centuries, but its existence would not be possible without the available data that represent the biological phenomena we wish to explore. The generation of one-dimensional sequence data (on which genomic research depends) began with the development of sequencing technology and the concepts that lie behind it.

2.2 GENOME SEQUENCING

Genome sequencing reduces a complex, three-dimensional DNA molecule into a linear, one-dimensional format, much like letters in a book except that there are no spaces or individual words, just a continuous stream of A's, C's, G's and T's. All we are left with is a sequential pattern of nucleotide bases, the majority of which, in bacteria, code for genes. The information in this pattern of bases, however, is informationally rich and determines the three-dimensional structure of proteins and various types of non-coding RNA molecules. Sequences involved in the rates of transcription and translation are also encoded in this one-dimensional representation and, importantly, the evolutionary history of the genome (and of individual genes) is recorded in the nucleotide pattern of DNA. When compared to homologous sequences from related strains and species, this pattern can elucidate phylogenetic relationships across groups of organisms and provide evidence of the effects of horizontal gene transfer by viruses and other mobile elements.

Cutting-edge sequencing technology is working towards the possibility of reading an entire genome in a single step, but over the short history of genome sequencing and even today a genome is sequenced as many separate fragments of DNA that are subsequently assembled and analysed. The associated biases, errors and practical limitations of this strategy have led to an explosion of ideas, software and technologies to handle the daunting task of genome sequencing and the complexity of the analyses that are needed to interpret the intricacies of life at a molecular level.

2.2.1 FIRST-GENERATION SEQUENCING

The first gene to be sequenced was that encoding the coat protein of the RNA bacteriophage MS2, a project carried out by Walter Fiers and colleagues in 1972 (Min Jou et al., 1972). This was followed four years later by the sequencing of the complete MS2 genome with its inventory of three genes (Fiers et al., 1976). The first generation of DNA sequencing technologies began with Sanger sequencing in 1977, an initiative led by Nobel Prize-winner Frederick Sanger. Sanger's sequencing method depends on a chain-termination step where deoxynucleosidetriphosphates (dNTPs) are added to a DNA template starting from a specific primer sequence using DNA polymerase. Di-dNTPs are also present in one of four reactions (one for each nucleotide) and, when incorporated into the growing DNA molecule, stop strand elongation due to the lack of a 3'-OH group required for bond formation between two nucleotides. The DNA fragments that result from rounds of strand elongation are heat-denatured and separated by gel electrophoresis according to size, a separate lane for each nucleotide (A, C, G and T). The bands of DNA are visualised using autoradiography and the sequence is determined directly from the gel image. This is possible because DNA fragments of different lengths all start from the same primer and a band appears where a ddNTP was incorporated into the sequence, meaning that consecutive bands represent consecutive nucleotides in the sequence (Sanger et al., 1992).

The timeline for the sequencing of larger and more complex genomes runs from bacteriophages to archaea to fruit flies, the nucleotide count increasing from thousands to millions and the simplistic organization of a single circular chromosome expanding into numerous linear ones. These early sequencing initiatives were completed using Sanger sequencing, each new project placing greater demands on the efficiency of the Sanger method. In particular, the Human Genome Project drove the modification of Sanger sequencing to become faster and more cost effective, leading to cheaper and more efficient protocols. Radioactive labels were replaced by base-specific fluorescent dyes and gel electrophoresis by automated capillary electrophoresis. Sanger sequencing remains the gold standard for clinical diagnostics, but it is too slow and expensive for most of today's studies unless they require the sequencing of only a handful of genes (Moorthie et al., 2011). The massive biological datasets that are routinely sequenced today are possible

because of a new generation of technologies: Next-generation sequencing (NGS), also known as high-throughput sequencing.

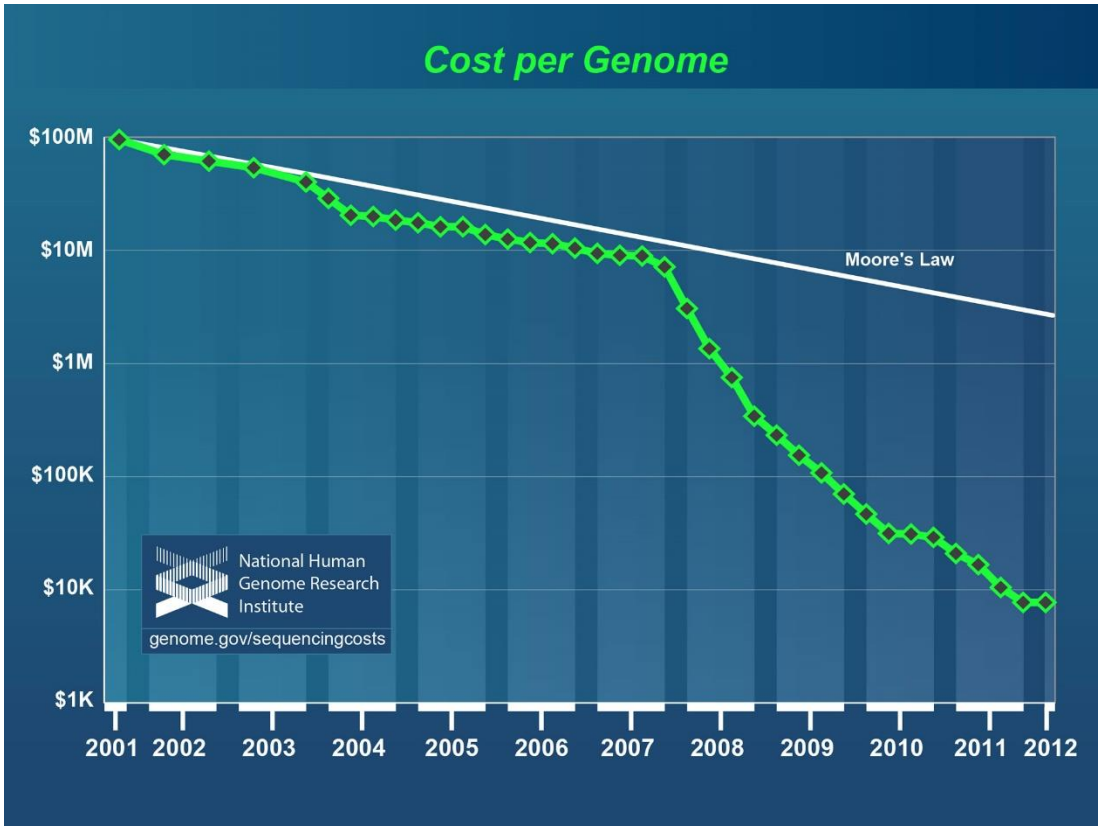


Figure 1.1: Cost of sequencing a genome over time. Reduction in the cost of sequencing the human genome over time has out-paced Moore's Law (Wetterstrand, 2012; Creative Commons license).

2.2.2 NEXT-GENERATION SEQUENCING

NGS technology is fuelled by the demand for cheaper and cheaper sequencing at ever higher capacities. Both Sanger and NGS sequencing rely on multiple copies of overlapping nucleotide 'reads' for the construction of longer sequences called contigs and for sequence validation, but Sanger creates one read at a time while NGS is massively parallelised. This is the defining characteristic of NGS sequencers - the ability to sequence millions or even billions of features together in a single run and to generate output where each feature is a sequenced read of nucleotide bases accompanied by per-base quality scores (Moorthie et al.,

2011). There are four main technologies that fall under the umbrella of NGS (as well as a host of minor ones) and they have had varying successes and contributions to the DNA sequencing revolution.

Pal Nyren and colleagues developed pyrosequencing, which depended on a luminescent method for measuring pyrophosphate synthesis (Ronaghi, 2001). This two-enzyme process used ATP sulfurylase to convert pyrophosphate into ATP that then acted as the substrate for luciferase, producing light proportional to the number of pyrophosphate molecules. The light intensity is measured as each nucleotide is washed over the template DNA and incorporated onto the growing strand by DNA polymerase. Sequence information can be observed in real time unlike the lengthy electrophoresis step required by Sanger sequencing and natural nucleotides are used instead of the modified dNTPs used in Sanger's chain-termination (Nyren, 2015). One issue with pyrosequencing is with the accurate detection of homopolymers due to the non-linear readout after four or five consecutive occurrences of the same nucleotide, making artificial insertion/deletion events a regular occurrence for genomic regions of low complexity (Balzer et al., 2011). This technology can produce reads of 400-500 base pairs (bp) and it increased sequencing speed by orders of magnitude compared to Sanger sequencing. Pyrosequencing was licensed to 454 Life Sciences and later purchased by Roche.

The Solexa sequencing method uses what has become known as 'bridge-amplification' to sequence both ends of a DNA molecule. This is achieved when replicating DNA strands adopt an arched configuration in order to undergo a second round of polymerisation off neighbouring surface-bound oligonucleotides. The first machines could only produce reads of up to 35 bp, but they had the advantage of providing paired-end data where the number of nucleotides that lie between two sequenced ends is also known (even if the nucleotides themselves are not). This additional information allowed for a more accurate determination of repetitive regions in a genome since the distance of two reads from each other could now be used to infer their relative positions. The first machine to use this technology was the Genome Analyzer, which was later followed by the HiSeq (longer reads and greater read depth) and then the MiSeq (lower through-put with longer reads than the HiSeq at a lower cost). Today, this technology is owned by Illumina, which has been by far the most successful sequencing company to the point of controlling most of the market (Metzker, 2010).

Applied Biosystems developed a method of sequencing by oligonucleotide ligation and detection (SOLiD). Unlike Sanger, 454 Pyrosequencing or Illumina, sequencing is carried out using a DNA ligase instead of a polymerase so the sequence-by-synthesis (SBS) technique that dominated innovation was absent from this technology. The company became Life Technologies after a merger with Invitrogen and while its shorter read length and lower read depth cannot compete with Illumina, its low cost has kept the technology in commercial use (Heather and Chain, 2016).

Life Technologies also developed the first sequencer to discard the use of detection by light for a 454-like protocol that measures nucleotide incorporation by the change in pH during polymerisation when protons are released. This technology allows for very rapid sequencing, but has the same limitation as 454 pyrosequencing when it comes to the reliable interpretation of homopolymers (Metzker, 2010).

The capabilities of sequencing technology have grown at a rate that has far outstripped Moore's law for computing, which predicts a doubling time of two years. For example, between 2004 and 2010 the capabilities of sequencing technology doubled every five months (Heather and Chain, 2016). This impressive rate of progress has led to a new generation of sequencers that are taking over from the Next-generation technologies.

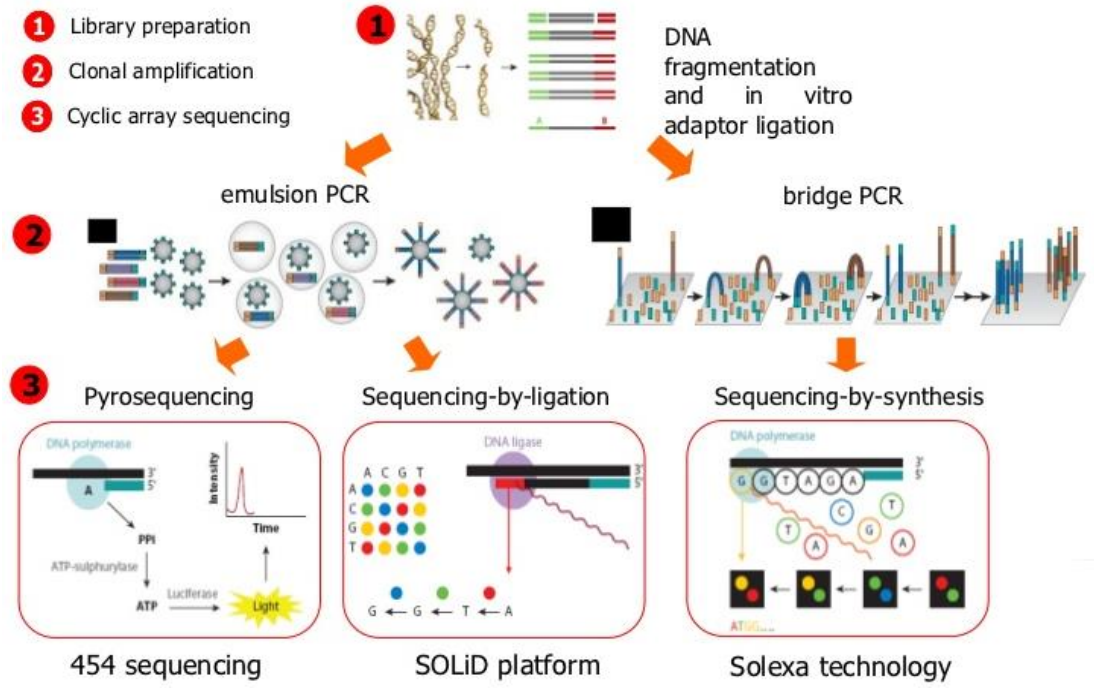


Figure 1.2: Next-generation sequencing. (Sanchez, 2011; CC Attribution-ShareAlike License)

2.2.3 THIRD-GENERATION SEQUENCING

Third-generation sequencers, although largely in development stages, emphasise the sequencing of single molecules without the need for DNA amplification. The biases and errors associated with DNA amplification are therefore absent from third-generation sequencers (Schadt et al., 2010). The most widely used technology at present is the SMRT (single molecule real time) platform from Pacific Biosciences. This technology can produce long reads up to 10 kb quickly, the sequencing of a single molecule occurring at the rate of the polymerase (Roberts et al., 2013).

A greatly anticipated third-generation technology is nanopore sequencing. Oxford Nanopore Technologies have already created a USB device the size of a mobile phone that was used to sequence the Ebola virus in Guinea by Joshua Quick and Nicholas Loman (Quick et al., 2016). Recent advances promise to take the sequencing monopoly away from a handful of commercial companies, giving small laboratories, research groups and even independent individuals the chance to

sequence the genomes of single organisms and of entire bacterial communities. Currently, however, error rates are very high when compared with Sanger and next-generation sequencing technologies and considerable improvements will have to be made before nanopore and related sequencing platforms gain widespread use (Lu et al., 2016).

It is likely that third-generation sequencing technologies such as nanopore will eventually output incredibly long reads with high accuracy, capturing a complete microbial genome in a small number of sequences. Bioinformaticians of the near future will take it for granted that the digital representations of genomes they analyse do not first have to be pieced together from a multitude of short reads and remain in a draft form, broken into numerous contigs.

Even after short reads are assembled together into larger portions of the genome, repetitive regions, limited read depth and sequence quality ensure that the genome remains in a draft form. Next-generation sequencing such as Illumina needs to be accompanied by gap-closure strategies such as manual PCR in order to turn a draft genome from numerous sub-sequences of varying sizes (or contigs) into a single sequence representing the entire genome. This is unfeasible for large projects containing dozens or even hundreds of microbial genomes, so genome analysis is often carried out on datasets of draft genomes. Alternatively, a complete reference genome can be used as a template for the mapping of reads, but this strategy also has its limitations because strain-specific genes are ignored as only genomic regions homologous with the reference are assembled.

The general problem of building genomes out of short reads is known as genome assembly and it is a procedure that is almost always completed before subsequent genome analysis takes place. As sequencing projects get more ambitious such as the goal of BGI (formerly Beijing Genomics Institute) to sequence one thousand genomes each from humans, microbes and animals/plants, the importance of sequence assemblers and their ability to handle multiple types of sequence data is growing.

2.3 GENOME ASSEMBLY

The first Sanger sequences were only a few dozen nucleotide bases in length, but these took weeks of laboratory work to produce and were remarkable

achievements for their time. The assembly of these early reads took no more than minutes, carried out largely by hand instead of requiring sophisticated software for the automated assembly of thousands to millions of short reads. As sequencers evolved and became more automated and more high-throughput, the demand for assembly software skyrocketed. The manual sequencing of increasingly longer DNA molecules was painstaking and it was accompanied by the impracticality of reconstructing more and more full-length sequences from reads by hand. Today sees the use of numerous popular assemblers and data from multiple sequencing technologies, leading to questions that several papers have already sought to answer: what combination of sequence technology and assembly software gives the most accurate representation of the underlying genome of interest, and how do we determine this optimal combination (Baker, 2012)?

2.3.1 PRE-ASSEMBLY QUALITY CONTROL

Sequencing data comes in the form of a multitude of fragmented, one-dimensional reads of A's, C's, G's and T's. The challenge is then to piece all these reads together to give one or multiple contigs of the genome sequence of interest. The quality of these reads will affect the quality of the final assembly so two important pre-assembly steps that are almost always carried out are read trimming and read filtering. Trimming removes low-quality portions at the beginnings or ends of reads and has been shown to increase assembly quality as well as the reliability of subsequent analyses (Del Fabbro et al., 2013). Numerous read-trimming software/packages/tools are available including Cutadapt, FASTX and Trimmomatic (Bolger et al., 2014). Read filtering is important for removing low-quality reads since their inclusion leads to a more fragmented and error-prone assembly. Tools such as Trimmomatic allow quality cut-offs to be specified so that the extent of trimming and filtering can be adjusted. The iterated adjustment of parameters is a common bioinformatic exercise since optimal parameter values are different depending on the structure and quality of data. In the case of quality control for sequencing data, a good strategy is to trim and filter reads at multiple cut-offs and use a tool like FastQC. This software assesses read quality as well as other factors like GC content, duplication level, k-mer content, read length distribution,

overrepresented sequences and the presence of adapters. Once sequence data have been appropriately trimmed and filtered, sequence assembly can begin.

2.3.2 ASSEMBLY ALGORITHMS

The assembler builds contiguous sequences by overlapping staggered reads, each nucleotide being represented multiple times. Longer reads like those from Sanger sequencing cover larger regions of a genome, have a higher probability of being unique portions of DNA and provide greater overlap with other reads of the same length. This is why Sanger sequencing is effective at low read depth (or coverage) while next-generation technologies such as Illumina and 454 Pyrosequencing require (and have the ability to generate) a lot more reads. Illumina also has the advantage of providing paired-end information by sequencing both ends of longer DNA fragments, ensuring that each read pair carries unique sequence information, especially since the insert sizes (the un-sequenced regions in the middle) of DNA fragments involved in paired-end sequencing are inexact and follow a distribution rather than a set value (Baker, 2012). What this means is that the probability of an identical read pair being sequenced twice is very low since varying insert size considerably increases the number of possible read pairs. This can be contrasted with single-end sequences, which have a higher probability of being sequenced more than once by chance alone. Duplication of reads due to PCR cycles is a separate problem and one that is usually resolved by software designed to identify and remove duplicates from sequence data (Ekblom and Wolf, 2014).

Assemblers designed to handle long-read sequences use an approach known as overlap-layout-consensus (OLC). Newbler, an assembler designed for 454 Pyrosequencing reads, uses this algorithm to assemble the relatively long reads generated by Roche sequencing technology. OLC is generally too computationally intensive for short-read data such as Illumina and SOLiD so alternative assembly strategies are available and can be separated into extension-based methods and De Bruijn graph algorithms. Extension-based methods are computationally efficient compared to OLC, but they are very sensitive to sequencing errors and repetitive regions. The De Bruijn graph algorithm is currently the most popular assembly method for short-read data and these assemblers dominate genomic research that uses Illumina (the majority) or SOLiD data (Miller et al., 2010).

The De Bruijn graph algorithm breaks each read into overlapping sub-sequences of a specified length, k, commonly referred to as k-mers. Each k-mer becomes a node in a network built by connecting all k-mers that overlap by k-1 bases. A contig is formed when overlapping k-mers reach a point that cannot be resolved by the assembler due to the unavailability of k-mers that extend the sequence (insufficient coverage) or due to the existence of repetitive regions that prevent the assembly algorithm from resolving k-mer positions. Paired-end information is used to bridge contigs in cases where each member of a read pair exists on a different contig. A string of N's then joins the contigs, representing the number of unidentified bases in a gap. The resulting structure is called a scaffold (Ekblom and Wolf, 2014). Popular De Bruijn graph assemblers include Velvet (Zerbino and Birney, 2008), SOAPdenovo (Luo et al., 2012) and SPAdes (Bankevich et al., 2012). The extension of contigs using short-reads and synteny information in the form of N's can be carried out using software such as GapFiller (Boetzer and Pirovano, 2012).

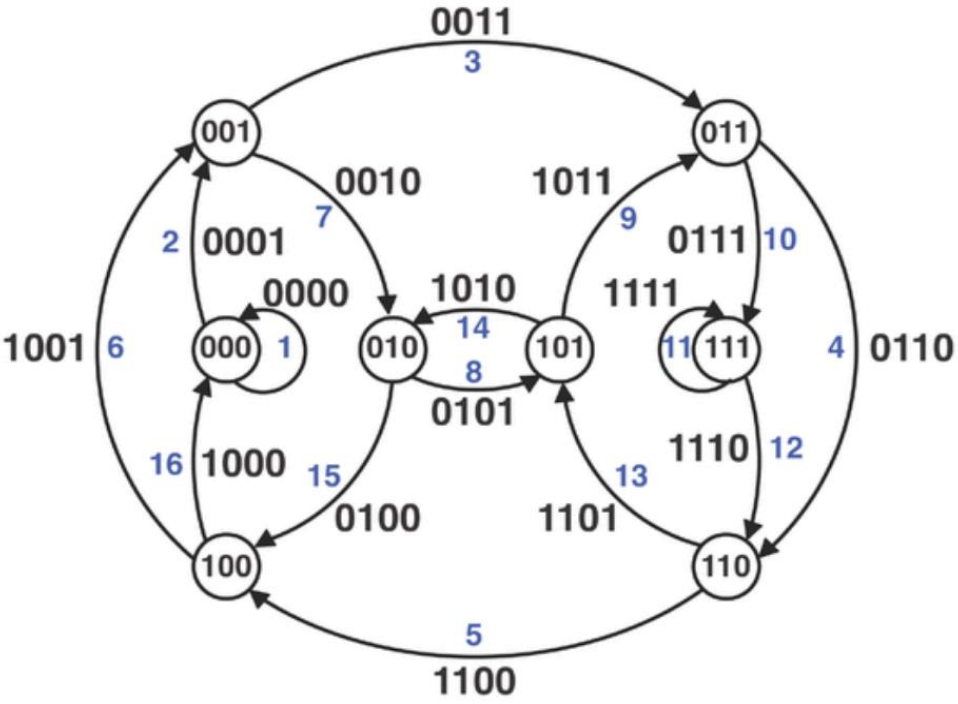


Figure 1.3: De Bruijn graph of the binary sequence 0000110010111101 using a k-mer size of four. The blue numbers from 1 to 16 trace the path of the sequence and arrows indicate the direction that the sequence moves through the path (Compeau et al., 2011; Springer Nature, License no: 4251410698290).

2.3.3 ASSEMBLY STATISTICS

A given assembly can be treated as a working hypothesis rather than an accurate representation of the sequenced genome. There are multiple parameters that need to be optimised during the assembly process including k-mer length, coverage cut-off for contigs and expected average coverage. A typical genome assembly process involves the adjustment of software parameters - and perhaps the testing of several different software - before an optimal assembly is chosen.

The assessment of assembly quality is a controversial issue because there is no consensus on what constitutes a good assembly. The sequence diversity across microbial genomes is staggering and considerable variation exists in all the following: genome size, replicon complexity, GC content, number and size of repeat regions, functional diversity and nucleotide k-mer composition. A small bacterial genome with no plasmids and few repeat regions will be far easier to assemble than one with multiple plasmids of different origin and whose genome is replete with repetitive genes.

Despite the challenges involved with assessing genome assemblies, numerous assembly statistics and other genomic analyses have been developed that are commonly used to measure assembly quality. The assembly statistics focus on the completeness of the assembly and the extent of read coverage. Assembly length (the sum of all contig lengths) is often compared to the length of complete genomes of the same species (if available) to test for agreement in size, which is usually very similar across strains of a species. N50 is the contig length at which 50% of the genome is contained in contigs of at least this length. The N50 value is a common assembly metric, but should be interpreted with caution as genomes with a greater number of repetitive regions will, on average, be more fragmented and have a lower N50. The largest contig size is another common metric reported with assembly summary statistics (Ekblom and Wolf, 2014).

The median coverage of contigs, nucleotides or k-mers (in the case of Velvet) is a useful statistic, but it obscures regional coverage variation over the length of an assembled genome. For this reason, numerous assemblers give a more detailed report of variation in coverage so that regions of low coverage (and hence of lower

confidence) can be detected. What constitutes sufficient coverage will vary however, depending very much on the genome of interest and the focus of the study (Sims et al., 2014).

Numerous other analyses are routinely carried out in addition to assembly statistics to test for genome completeness and lack of contamination. The absence of essential genes is evidence of missing genome regions caused by insufficient coverage or quality issues. Coding and non-coding genes involved in transcription and translation can be used as marker genes to test assembly quality. If genes essential for cell viability are absent from an assembly, confidence in the presence/absence distribution of other genes in the genome is low. Multiple steps in the annotation of a genome are also used to test overall assembly quality and lack of contamination. Comparative Genomics involving strains of the same or different species can highlight potential issues such as missing genome regions, contradictory genome statistics (GC content, genome size, etc.) and a high percentage of genes with top hits to other species (usually a sign of contamination rather than horizontal gene transfer).

High-quality assemblies are the foundation on which reliable bioinformatic analyses are based. Without detailed annotation however, a genome is just a sequence of unintelligible nucleotide bases, one after the other. Looking at an assembled genome on a computer screen, it is impossible to tell what species it encodes or whether the genome belongs to an animal or a plant or a bacterium. The hidden patterns inherent in a genome's nucleotide order need more sophisticated methods to be deciphered. The expanding array of techniques and software currently being employed in genomic and comparative genomic studies is unravelling the mysteries of DNA, elucidating the functional and phylogenetic properties of organisms on a molecular level.

2.4 GENOME ANNOTATION

Genome annotation has tedious connotations. It conjures up the manual curation of open reading frames (ORFs) in software such as Artemis – a tool for sequence visualisation and annotation (Rutherford et al., 2000). Annotating a genome however, is what extracts structural and functional information from the organism of interest. It describes the physiological capabilities of a microbe and

informs decisions on whether the microbe is pathogenic, probiotic or of industrial or commercial use.

The exponential increase in the number of sequenced genomes and the growing sample size of comparative genomic studies have forced annotation tools to become faster, more automated and more accessible. While manual annotation tools such as Artemis work on a gene-by-gene basis, gene prediction software and functional databases can be combined to annotate hundreds of genomes in anywhere from several hours to a few minutes (depending on the speed of software and the size of the database).

2.4.1 GENE PREDICTION

The average protein-coding content of a bacterial genome estimated from 2,671 Genbank genomes in 2014 is 88% (Land et al., 2015). The majority of a bacterial genome encodes the proteins necessary for survival and reproduction in a given environment so an essential step in genome annotation is to predict the gene content of a given genome. This can be done with homologous gene searches using curated gene databases and sequence alignment software such as BLAST (Altschul et al., 1990). The exponential increase in the number of new genomes being sequenced (and hence the number of novel genes) means that genome annotation using known genes is often insufficient, leading to the necessity for *de novo* methods that do not rely on reference sequences. There are multiple technical and conceptual issues that must be overcome in order to accurately predict the set of genes that a genome codes for using *de novo* methods. There are numerous important factors to consider when predicting genes in a microbial genome: gene length, k-mer content, GC content, relative gene positions, gene overlap, correct start-site prediction and reading frame, among others. These factors rest on the concept of an open reading frame (ORF), a basic genetic principle that must be incorporated into every gene prediction software and every study involving genome annotation.

A genome may appear to be a random ordering of four nucleotide bases, but it has a structure defined by its molecular transcription and translation machinery that goes beyond the genetic code of triplet codons representing amino acids. A gene can lie on either strand of a DNA double helix and because of the reverse

complementarity of DNA, the opposite strand also contains the gene sequences contained on the other, but the nucleotides are in the reverse orientation and complemented according to the rules of base pairing (Alberts, 2015). A simplistic model of a genome would thus have two reading frames, forward and reverse, and RNA polymerase would always transcribe genes in one of two frames, passing mRNA to the ribosome and its associated molecules, which translate the gene three bases at a time, one codon after the other. Stagger an inter-genic region by inserting or deleting a nucleotide however, and the reading frame of downstream (3' direction) genes is altered relative to those upstream (5' direction). Mutations involving the insertion or deletion of nucleotides need to involve multiples of three nucleotide bases to preserve the reading frame relative to downstream genes. This means that there are six potential reading frames for genes on a double-stranded DNA molecule and a gene can lie in any one of these. An open reading frame is therefore a sequence of nucleotides whose length is a multiple of three, that begins with a start codon (ATG, TTG or GTG coding for methionine) and ends with a stop codon (TAG, TGA or TAA, signalling to the ribosome to terminate translation) where no other stop codon appears in the interval of codons (Cristianini and Hahn, 2007).

An ORF is the minimal requirement for the prediction of a complete, functional gene. It would be relatively easy to predict all ORFs in a genome and then impose a method of filtering false positives by using gene length and overlapping sequences (for example, excluding a small ORF within a larger ORF on the same or different reading frames). All ORFs are not necessarily genes however, and many start codon 3-mers can appear within a gene (just as many methionine residues can exist within a protein), making accurate start-site prediction difficult. Choosing the longest ORF is often dangerous because genes in different reading frames can exist upstream of the correct start codon, excluded due to the incorrect upstream extension of a gene to maximise ORF length. An effective solution to this issue is the inclusion of information related to nucleotide composition in gene prediction algorithms.

The ability to differentiate between coding and non-coding sequences of a genome is the key to a successful gene prediction algorithm. The nucleotide composition of coding DNA is substantially different from non-coding DNA due to selective constraints and other factors (Hayes and Borodovsky, 1998). A summary of the average nucleotide composition between a gene and the surrounding inter-genic region in terms of k-mer content (for example) can show obvious differences.

Predicting the transition from coding to non-coding sequence, or vice versa, is more challenging however, because the correct codon (either start or stop) has to be identified that divides the two regions. For start-site prediction, it is always possible to extend an ORF upstream to the next start codon in the same reading frame – the challenge here is to determine which start site is more probable. The fact that gene prediction must take place over six reading frames only complicates matters, and numerous software-mediated approaches refine initial gene predictions by considering the relative position of other genes within the genome or using sequence patterns from microbial ribosomal binding sites (RBS) of known bacteria (Lukashin and Borodovsky, 1998). The existence of introns in eukaryotes is a separate issue and one that can be ignored when considering prokaryotic gene prediction software (Wang et al., 2004).

Markov models have proven to be an effective tool in differentiating coding from non-coding sequences and therefore providing accurate predictions of genes and their start sites (Cristianini and Hahn, 2007). These models make the assumption that a state (in this case, the occurrence of a nucleotide) depends only on a specified number of previous states in the model (the nucleotides immediately upstream). A first-order model is one that depends only on the previous state while a second-order model depends on the two previous states, and so on. Models of this kind can be expanded into a hidden Markov model (HMM) where a sequence is composed of unobserved (or hidden) states, transitioning from one state to another along the sequence length and the observed nucleotide pattern follows the probabilities of these hidden states. HMMs used in the prediction of genes have multiple states, including the coding and non-coding regions of the genome as well as start/stop codons and reverse complemented sequences on the opposite strand.

A common strategy in gene prediction software is to first identify all possible ORFs in a genome's six reading frames and then to apply Markov models in order to predict the subset of ORFs that represent coding regions of the genome. GeneMark.HMM (Lukashin and Borodovsky, 1998) adopts this strategy and uses nine HMM states to differentiate between coding and non-coding sequences. This software uses a fixed-order Markov model, which means that it uses a pre-determined number of states (nucleotides) to predict a subsequent state. In this case, second-order models are used so that a nucleotide's identity depends on the previous

two nucleotides in a sequence. Gene start-sites are refined in a subsequent step that uses an model to identify upstream ribosomal binding sites (RBS).

Glimmer3 (Salzberg et al., 1998) adopts a different strategy, predicting an initial set of genes and using the derived k-mer counts to build models that are then used to iteratively predict genes with greater accuracy. It uses what are called interpolated Markov models (IMM) to vary the order of its Markov models depending on available data. This is an alternative approach to the fixed-order models of GeneMark.HMM because higher-order models are used when sufficient k-mer counts are available and lower-order models are used in cases where accurate probabilities for higher-order models cannot be estimated. The initial step of Glimmer3 is to identify all possible ORFs, but its Markov models are more flexible because higher-order models decrease the accuracy of gene prediction when the k-mers they use to predict a subsequent nucleotide are not sufficiently represented in a genome. What constitutes sufficient representation of k-mers is an issue that highlights a more general challenge in genomic analysis: there is usually no correct set of values for a given parameter (such as order in Markov models) and adjustment of parameters while measuring prediction accuracy is often the best approach when deciding on the most appropriate values to use.

The prediction of partial genes is important when draft genomes are being studied because genes can be truncated at either their 5'- or 3'-end (or both) at contig boundaries. The reason for the existence of partial genes in draft genomes reflects the inability of assembly algorithms to resolve repetitive regions. Genes can be present in multiple copies (such as the 16S rRNA gene) or can have highly conserved domains as part of a larger gene family. Both cases introduce repetitive regions into a genome that an assembler will fail to assemble fully. Alternative explanations for partial genes are low read coverage and low read quality, but these issues can be avoided with competent pre-sequencing and sequencing steps.

GeneMark.HMM allows for the prediction of partial genes while Glimmer3 does not. MetaGene (Noguchi et al., 2006) predicts complete and partial genes in genomic and metagenomic microbial datasets. It uses GC-dependent di-codon frequencies from bacterial and archaeal species along with numerous ORF statistics to assign scores to pre-computed ORFs. The statistics are calculated from the input (meta)genome and involve numerous distributions including ORF length, distance of

predicted start-sites from annotated start-sites and orientation-dependent distances of neighbouring ORFs.

The three gene-prediction software described above have parallels in other available software and there are dozens of downloadable programs and web tools dedicated to the prediction of prokaryotic genes. The most popular software have high sensitivities and specificities, but no gene prediction strategy is perfect and each method has its strengths and weaknesses. A strategy that is commonly used to reduce false negative gene predictions is to combine the results of multiple software and include all predicted genes in the final output (or perhaps a gene predicted by two out of three software). An accurate set of predicted genes is essential for subsequent genome annotation and analysis. This is also true for the prediction of non-coding genes such as tRNA and rRNA sequences, which have equivalent dedicated software such as tRNAscan-SE (Lowe and Eddy, 1997) and RNAmmer (Lagesen et al., 2007). Subsequent annotation steps are equally important, involving the use of appropriate gene databases and the assignment of function to predicted genes based on homology.

2.4.2 FUNCTIONAL ANNOTATION

SWISS-PROT is a high-quality, curated, protein sequence database that maintains a high level of integration with other databases (Bairoch and Apweiler, 2000). A query gene (usually the translated amino acid sequence) can be BLASTed against this database and hit a sequence with 100% identity over its full length, indicating that the function of the SWISS-PROT sequence can be reliably transferred to the query gene. Alternatively, a BLAST result can give back lower alignment statistics that typically cover a range of values, which makes assigning functions to genes an issue of choosing parameter thresholds (i.e. what is the lowest percentage identity and alignment length at which a function can be transferred from a reference to a query gene?) If the top BLAST hit for a query gene falls below chosen thresholds, the gene is labelled as ‘hypothetical’ unless another database or method is able to assign functional information. For instance, the query gene can be BLASTed against the much larger, NCBI non-redundant (nr) database of protein sequences – a database that receives a much lower level of curation than SWISS-

PROT (although SWISS-PROT is a subset of NCBI nr). The query gene might then hit a reference sequence that passes the previously chosen BLAST thresholds and be assigned the function of the reference sequence based on the assumption that the thresholds indicate a homologous relationship of sufficient similarity that a shared function can be inferred. Failure to assign a function to a query gene using BLAST (or related local and global alignment software such as USEARCH (Edgar, 2010) is often remedied using HMM databases such as TIGRFAM (Haft et al., 2003) and Pfam (Finn et al., 2016) that capture conserved domain information and have the ability to detect more distant homology. A hypothetical gene has, of course, a function (if it is real and not a false positive gene) – it is just that no homologous sequence has been identified in a database from which a function can be inferred.

The above common scenario highlights the interplay between databases, algorithms and parameter thresholds when assigning functions to predicted genes in newly sequenced genomes. Just like the careful selection of gene prediction software and appropriate parameter settings lead to an optimal set of predicted genes, so too do the choice of database, algorithm and thresholds lead to the most accurate assignment of functions. These choices can vary, of course, depending on the desired outcomes of an analysis, especially whether the study design aims at giving a general overview or focusses on very specific aspects of certain genes and functions.

The Clusters of Orthologous Groups (COG) database (Tatusov et al., 2000) assigns genes from numerous species to a hierarchy of nested functional groups, the highest of which symbolises general functions by individual letters (i.e. 'G' for 'Carbohydrate transport and metabolism'). In this case, each query gene is assigned to a COG letter based on relatively lenient BLAST thresholds (i.e. 40% identity and 50% of query gene aligned) and those genes that fall below cut-off values are interpreted as hypothetical genes. Interestingly, COG has two general categories - 'General functional prediction only' and 'Function unknown' - that are equivalent to the category 'hypothetical' at this functional level. These categories highlight a larger issue in other databases of the presence of sequences that have 'hypothetical' as their only functional annotation. This issue is particularly prevalent in the 2017 version of the NCBI non-redundant protein database where many functionally annotated sequences are identical to 'hypothetical' sequences from separate projects. The top hit for a query gene can therefore be a hypothetical protein at 100% identity over its full length while the second hit can be just (or almost) as good, but involve

an assignable function. An obvious method to combat this problem is to exclude all hypothetical proteins from a database before using it as a reference, since predicted genes that fall below specified thresholds can be labelled as 'hypothetical' anyway.

It is possible for a gene to be assigned multiple COG letters, which emphasises a fact that is often overlooked: a single protein can have multiple domains, each one with a very different sub-function, together defining the function (or functions) that the protein carries out as a whole. Multi-domain proteins are more prevalent in eukaryotes (Jacob et al., 2007), but there is evidence to suggest that up to two-thirds of prokaryote proteins have at least two domains, suggesting that functional annotation of microbial proteins should be restricted to individual domains (Vogel et al., 2004). This is where HMMs become more useful than general BLAST searches. A hidden Markov model can be created from many homologous sequences, representing the sequence variation within a single domain. Amino acid sequences are usually used for building HMMs because they can detect more distant homology, but nucleotide sequences can also be used and are the input data for building models such as those involved in the gene prediction algorithms described earlier. A large database of domain-centric HMMs is a powerful tool for predicting the functional variation of a gene set, focussing on conserved, functional sequences within genes rather than treating the entire sequence as an entity that must share homology over its full length. BLAST specialises in the local alignment of exact sequences, but HMM searches require a different algorithm like that incorporated by hmmscan from the HMMER3 suite of tools. This software uses profile HMMs built from multiple alignments of conserved domains stored in databases like TIGRFAM and Pfam mentioned earlier. The output is a BLAST-like report of statistics that reflect the probability of regions of input query genes sharing domains contained in the profile HMMs. Analyses like these (as well as BLAST) can be highly parallelised to run using many HMMs on a multitude of query genes.

Strategies involving the use of BLAST or HMMER on input sequences predict the presence (or absence) of particular genes and domains. Subsequent annotation steps can collect these isolated results together and present them in an interactive framework where they reveal the biological context of the genes and genomes being studied. The variety of ways this can be done is tremendous, depending on the scope of individual projects and research questions. Several software tools have been developed to present gene annotation results in the larger

context of the biochemical pathways of which they are a part. The Kyoto Encyclopedia of Genes and Genomes (KEGG) displays this information as graphical diagrams of metabolic and regulatory pathways (Ogata et al., 1999). This and related methods of visualising networks of interacting proteins can reveal patterns that lists of individually annotated genes cannot because the functioning of entire pathways is considered instead of the presence or absence of single functions in isolation.

A comprehensively annotated genome provides a powerful insight into the inner workings of a microbial strain. There are numerous complete genomes on GenBank that represent the type strains of species and they have been downloaded many times. Along with a file holding the unbroken string of nucleotides that represents all the strain's genetic information (excluding potential plasmids, which are stored in separate files), an array of additional files is also available that encompass the genome's functional repertoire. As insightful as this information can be, a type strain does not reflect the variation within a species – it merely acts as a definition of the species main genotypic and phenotypic characteristics. The effects of gene loss and horizontal gene transfer (HGT) mean that every gene within a species has a distribution that falls somewhere along a range from being universally present (a core gene) to being restricted to a single genome (strain-specific). Even more importantly, type species of genera and sub-genera fail to capture the enormous diversity of functions within their clades, especially for paraphyletic ones such as *Lactobacillus*, which was shown to contain multiple other genera branching from within its phylogenetic tree (Salvetti et al., 2013).

This is where Microbial Genomics gives way to Comparative Microbial Genomics, an extension of assembly and annotation procedures to multiple genomes followed by comparative genomic techniques. Describing the genomic details of a single strain is like taking a snapshot of a dynamic process and hoping to understand the forces involved without measuring how these forces change the underlying biology over time. Comparative Genomics applied to *Lactobacillus* (the subject of the rest of this review) takes the genomic information from many separate strains, combining them to reveal the evolutionary and ecological processes that explain the diversity that lies behind this commercially important genus.

3 COMPARATIVE GENOMIC STUDIES AND EMERGING CONCEPTS

3.1 EARLY COMPARATIVE GENOMIC STUDIES IN LACTOBACILLI

The first comparisons involving sequences pre-dated whole-genome comparative studies of cellular microbes. Single-gene comparisons expanded into analyses of the homology and synteny of prophages, showing that these diverse replicators can share similarities in gene sequence and gene order (Koonin & Galperin, 2003). A scan of PubMed shows that early microbial studies of the late 1990s and early 2000s are dominated by the comparison of prophages, a trend that is as true for *Lactobacillus* as it is for other genera. Desiere *et al* describe sequence similarity and synteny in the late gene cluster of *Lactobacillus* phages (Desiere et al., 2000). Comparison of complete prophages highlights problems with phage taxonomy in lactic acid bacteria (Proux et al., 2002), describes prophage diversity across *Lactobacillus* strains (Ventura et al., 2003, Ventura et al., 2004) and across species (Tuohimaa et al., 2006), and provides insight into the inconsistent phylogeny of prophages with their bacterial hosts (Ventura et al., 2006). Comparative genomics of prophages reveals in greater detail the mosaic nature of viral DNA, capturing the rapid rate at which phages evolve and diversify over time, even within closely related host strains and species.

The intervening years have seen an increasing number of comparative genomic studies based on bacterial genome sequences. Comparative genomics does not always involve the analysis of sequence data, however. A study of *Lactobacillus sakei* strains isolated from meat used comparative genome hybridisation (CGH) to compare 18 strains with a reference *L. sakei* strain, 23K (Nyquist et al., 2011). These methods rely on hybridisation to known sequences and were the method of choice for these types of analyses before direct sequencing of DNA became fast and affordable. Today, the analysis of genome sequence data is becoming more common place and a lot more high-throughput as these trends continue at an ever increasing rate.

Makarova & Koonin carried out an early comparative genomic study on lactic acid bacteria involving nine genomes, emphasising the phylogenetic and

functional diversity of LAB organisms (Makarova et al., 2006). They showed that the nutritionally rich environments typical for LAB species have apparently selected for considerable gene loss across the group, their genomes displaying a limited range of biosynthetic capabilities. These varying levels of auxotrophy are balanced by a broad range of carbon and nitrogen transporters as well as key HGT events that have allowed LAB species to successfully adapt to these habitats. The study also highlights the paraphyletic nature of *Lactobacillus*, a genus that has five other genera branching from within its phylogeny (Makarova et al., 2006).

Makarova & Koonin focus on many of the important concepts in Comparative Microbial Genomics, concepts that will act as a template for the rest of this review. Selection pressure, gene loss, horizontal gene transfer, niche-specific adaptation, strain-specific genes and paraphyletic genera – all these phenomena are essential for understanding the evolution of *Lactobacillus*, its current paraphyletic status and the phylogenetic and functional diversity that characterise its members. A good place to start when thinking about these issues is the concept of the pan-genome, the intriguing fact that the total number of non-homologous genes within a collection of microbial strains of a species can far outnumber the gene count in any one of their genomes. The pan-genome concept is generally applied to an individual species, but it can be extended to multi-species comparisons, a topic that will be discussed in later sections.

3.2 THE PAN-GENOME

3.2.1 HOMOLOGY, ORTHOLOGY AND PARALOGY

For the pan-genome of a species to be described or, indeed, for it to make sense at all, some evolutionary concepts need to be taken into consideration. The sequences of a gene present in multiple strains of a species need to be identified as having a common ancestor: a gene present in a single copy in one bacterial cell before subsequent rounds of replication led to the diversification of strain lineages. The ancestral gene no longer exists, of course, but the conserved and variable regions of its descendant genes represent the evolutionary trajectory of the gene as it diversified over time.

The modern sequence descendants of an ancestral gene are referred to as homologous genes – a central concept in the comparison of all genomes, both

prokaryotic and eukaryotic. In Comparative Genomics, an efficient way to identify homologous genes across genomes is by using bi-directional best hits (BBH) (Ward and Moreno-Hagelsieb, 2014). In this method, all genes from one genome are blasted against all genes from the other and a gene pair that have each other as their respective top BLAST hits are called BBHs and assumed to be homologs.

The accuracy of the BBH method depends on appropriate parameter cut-offs such as percentage identity and alignment length. It is favourable to use additional steps to support this initial prediction of homology and this is what software like QuartetS (Yu et al., 2011) and OrthoMCL (Li et al., 2003) do. They use Markov clustering to group homologous pairs of genes, defining a gene across all genomes in a dataset by the BBH pairs that cluster together. For instance, a gene cluster that has a sequence present in every genome represents a core gene.

There is an additional complication, however. The sequences in a homologous gene pair are either orthologs or paralogs of each other. Orthologous genes exist in different genomes and arise from a common gene ancestor through binary fission of the parent cell (and its accompanying DNA) into two daughter cells. Paralogous genes, by contrast, can exist either in the same genome or different genomes, arising from a gene duplication event. It is no contradiction that a duplication event can lead to a paralogous gene pair in separate genomes; gene duplication produces gene B from gene A and replication produces two orthologous copies of these genes in another cell. Gene A is therefore an ortholog of gene A and a paralog of gene B in the other cell.

The problem arises when the loss of gene A in one species and the loss of gene B in another leads to the identification of gene B as a BBH with gene A from the other species, a homologous relationship that would then be mistaken for orthology. QuartetS deals with this problem by constructing quartet gene trees composed of a BBH pair and an identified BBH paralogous pair from a third genome of the same gene cluster. If the split between the paralogous genes occurs first, followed by a subsequent split that creates the homologous BBH pair of uncertain origin, then the pair are assumed to be paralogous, mirroring the relationship of the paralogous BBH pair that gave rise to them (Yu et al., 2011).

When considering the pan-genome - the total number of genes in a genomic dataset (usually confined to a single species) of a given size - the main point of interest is not the total number of sequences; it is the total number of genes (grouped

into orthologous clusters) along with their distribution across the strains under study. The pan-genome includes genes of potential orthology with other sequences outside of the dataset, but which have no identified homology with any genes in the genomes under study. These genes are called strain-specific genes, unique genes or 'orphans' and will be covered in more detail in a later section.

3.2.2 THE EUKARYOTIC PAN-GENOME

In many eukaryotic species, the genes present in a genome remain the same from organism to organism. The main source of genetic variation resides in homologous sequences that differ in their nucleotide and/or amino acid composition in the form of single nucleotide polymorphisms (SNPs) or, more uncommonly, insertion/deletion events. Non-homologous genetic differences such as copy-number variation (CNV) involve the deletion or duplication of repetitive regions and also play a role in eukaryotic genetic variation, although CNV has also been detected and studied in prokaryotes (Taniguchi et al., 2010).

Over the past number of years, a growing appreciation for the existence of eukaryotic pan-genomes has swept through the scientific community following a number of key studies. The phytoplankton, *Emiliani*, shows considerable variation in gene content over a broad geographical area with only two-thirds of the pan-genome shared by all sequenced isolates. Genes coding for metal-binding proteins, in particular, display variable presence and help explain this species' physiological plasticity over different aquatic environments (Read et al., 2013). Plant genomes, too, show variation in gene content (Hirsch et al., 2014) and studies are now adopting the concept of the pan-genome for research in human genetic variation (Li et al., 2010). Variable gene content within species across the tree of life is increasingly seen as a driving force for phenotypic variation, contributing to explanations of how organisms within a species adapt to particular niches. Nowhere is this more extensive than in prokaryotes and the intra-species diversity within *Lactobacillus* offers many examples of how a pan-genome expands the environmental range of a species. This allows them to thrive in niches that would be inaccessible if bacterial evolution depended solely on the environmental selection pressure placed on mutations in homologous DNA.

3.2.3 THE PROKARYOTIC PAN-GENOME

The concept of the pan-genome was developed by Tettelin *et al* in 2005 (Tettelin et al., 2005). In this study they sequenced the genomes of six strains of *Streptococcus agalactiae* and showed that the gene content across the strains could be divided into three groups: genes shared by all strains (the core genome), genes shared by a subset of strains (the accessory genome) and genes unique to each individual strain (strain-specific genes). The pan-genome can be visualised as a curve on a graph of number of genomes (x-axis) versus number of genes (y-axis), the addition of each new genome adding extra genes that have no close homology with those of the previously added genomes. In this way, the pan-genome for a given genomic dataset consists of all core genes plus dispensable and unique genes. Tettelin *et al* extrapolated this pan-genome curve forward and hypothesised that the appearance of new genes would continue even after the addition of hundreds of genomes (Tettelin et al., 2005). In contrast, 44 genomes of *Streptococcus pneumoniae* were used to predict that the pan-genome would become saturated at 50 genomes with no new genes added following further genome addition (Donati et al., 2010).

Just like the pan-genome curve, core-gene and new-gene curves can also be plotted, representing the tendency of genetic diversity to be more comprehensively captured with larger genomic datasets than smaller ones. The pan-genome of a particular species (for a given number of genomes) is said to be either open or closed, depending on whether the addition of more genomes will continue to introduce “new” genes into the dataset. The calculation of this value is based on the slope of the log of new genes plotted against the log of number of genomes; if this value (α) is less than one, the pan-genome is open, otherwise it is closed (Tettelin et al., 2005). An open pan-genome is a useful indicator of the extent of genetic diversity within a microbial species, suggesting frequent HGT and gene loss events, but it can also be over-interpreted. The openness of a pan-genome depends very much on the size of the dataset and, in principle, even the most diverse species will have a closed pan-genome when enough genomes are added to the dataset under study.

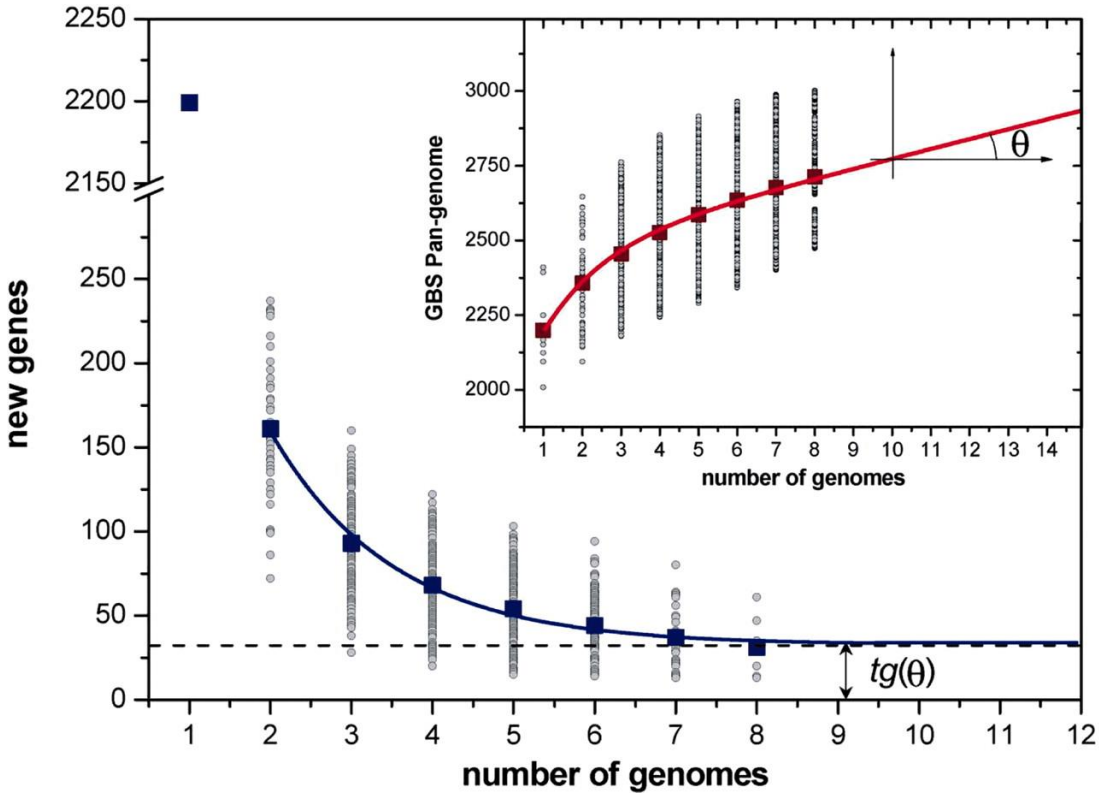


Figure 1.4: New- and pan-genome curves of *Streptococcus agalactiae*. New and total genes are plotted as a function of number of genomes and the order of addition of genomes is permuted 1,000 times to give measures of variation (circles) and average values (squares) for new genes and total genes, respectively (Tettelin et al., 2005; Copyright (2005) National Academy of Sciences).

3.2.4 THE PAN-GENOMES OF SPECIES WITHIN THE *L. CASEI* GROUP

Broadbent *et al* analysed the genomes of 17 strains of *L. casei* isolated from dairy, plant and human sources (Broadbent et al., 2012). This dataset had an open pan-genome with 1,715 core and 4,220 accessory genes. They estimated that the pan-genome was 3.2 times larger than the average size of individual genomes, suggesting frequent HGT from other lactobacilli and more distant bacteria. Dairy strains displayed considerable gene decay, hypothesised to be due to relaxed selection pressure in nutritionally rich dairy environments (Broadbent et al., 2012).

Smokvina *et al* analysed 34 strains of the closely related *L. paracasei*, showing that each genome had between 2,800 and 3,100 protein-coding genes with a conserved species core of 1,800 and a pan-genome of 4,200 genes (Smokvina et al., 2013). Accessory genes mainly consisted of hypothetical proteins, phages, plasmids and transposases as well as those functional groups that are known to undergo niche-dependent selection pressures such as transporters, CRISPR-associated proteins, EPS biosynthesis proteins and cell-surface proteins. An enormous variety of sugar-utilisation gene cassettes reflected the adaptability of *L. paracasei* to different niches with strains harbouring between 25 and 53 cassettes. Despite this, no obvious relationship was found linking gene content to niche, highlighting the complex evolutionary relationship that this species has with its environment (Smokvina et al., 2013).

The studies of Broadbent and Smokvina are largely consistent, describing an open pan-genome with considerable variation from strain to strain. Douillard *et al* conducted a reference-based study of the third member of the *L. casei* group, *L. rhamnosus*, using 100 strains and the widely used probiotic *L. rhamnosus* GG (Douillard et al., 2013b). They predicted 17 highly variable genomic regions related to lifestyle and showed that phylogeny could be partly associated with niche. Unlike the outcomes of the comparative studies of *L. casei* and *L. paracasei* described above, variation in gene content was limited to genes that were present in *L. rhamnsus* GG. This strategy has the advantage of defining inter-strain variability in terms of a complete, well-annotated reference genome that has been comprehensively researched, but suffers from the exclusion of genes and genomic regions that are absent from strain GG. As such, the study does not strictly deal with the pan-genome of *L. rhamnosus*, but it does highlight the important point that a reference strain fails to capture the extent of the functionality within a species. Kant *et al*, however, did describe an open pan-genome for *L. rhamnosus* in their study of cell-surface proteins (Kant et al., 2014), showing that a focus on strain-specific properties highlights the genetic diversity of this species.

The *L. casei* group have been described as niche generalists that have adapted to very different habitats (Cai et al., 2009). Add to this the frequent taxonomic inconsistencies that have led to numerous strains being misclassified within the *L. casei* group (Wuyts et al., 2017) and an open pan-genome might seem like an obvious comparative genomic conclusion. Depending on the study/species, this

might be because of frequent HGT and gene loss across diverse habitats or due to exaggerated diversity within species introduced by incorrectly classified strains. An interesting question to ask is if the apparently open pan-genome of *L. casei* exists for other *Lactobacillus* species, perhaps those with more reliable taxonomy or those with smaller genomes and a narrower range of habitats.

3.2.5 THE PAN-GENOMES OF OTHER *LACTOBACILLUS* SPECIES

Martino *et al* analysed the genomes of 54 strains of *L. plantarum* isolated from different environments (Martino et al., 2016). *L. plantarum* is a generalist species with a large genome and, like species from the *L. casei* clade, strains often harbour more than 3,000 genes, which is a large gene count for a species of *Lactobacillus*. The study labelled *L. plantarum* as a nomadic species and used gene-trait matching to show that strains do not cluster according to their source of isolation. They revealed a mixed distribution of strains where the phylogeny and function did not explain adaptation of groups of *L. plantarum* to specific environments (Martino et al., 2016). This is a good example of a generalist strategy where potential niche-specific adaptations are not found to be exclusive to strains isolated from the niche of interest. Martino *at al* describe the *L. plantarum* pan-genome as not having reached saturation at a sample size of 54 and it can be hypothesised that this species, like the generalist *L. casei*, also has an open pan-genome.

Ojala *et al* described the core and pan-genome of 10 *L. crispatus* strains (Ojala et al., 2014). Strains of this species have smaller genomes than *L. plantarum* and the *L. casei* group, but it would not be considered a specialist. While *L. crispatus* is commonly isolated from the human vagina, it is also found in a variety of other host-associated habitats. With a core genome of 1,224 genes and an accessory genome of 2,705 genes, these 10 strains are predicted to have an open pan-genome that continues to rise with the addition of new genomes until a sample size of 285 genomes has been reached (Ojala et al., 2014).

Frese *et al* showed that different *L. reuteri* lineages have become adapted to living in the gut of their respective vertebrate hosts, clustering together based on multi-locus sequence analysis (MLSA). Strains isolated from rodents show a large,

adaptable open pan-genome while strains isolated from humans have undergone greater genome reduction and reveal a closed pan-genome (Frese et al., 2011). These results are a better example of niche-specific adaptation, highlighting the application of the pan-genome concept within sub-lineages of a species, a useful strategy when gene flow and host specificity are the phenomena of interest. Wegmann *et al* carried out a core genome alignment of 20 *L. reuteri* genomes, showing that strains isolated from the same vertebrate tend to cluster together as sub-clades, but do not always represent a monophyletic group (Wegmann et al., 2015). The study focussed on pig isolates and described 6 strains having a core genome of 1,364 genes and a pan-genome of 3,373 genes. The core genome decreased to 851 genes and the pan-genome increased to 5,225 genes when the 14 strains from other hosts were included, demonstrating the reasonable point that more distantly related strains will likely share fewer genes and possess a more variable accessory genome.

L. reuteri has also been isolated from non-intestinal environments such as sourdough (Zheng et al., 2015) and its niche-dependent description as having both an open and a closed pan-genome shows that the pan-genome concept is very much an analytical tool rather than a factual description of a species that takes one of two values, true or false, open or closed.

Kant *et al* expanded pan-genome analysis to 20 complete genomes from across the *Lactobacillus* genus. The pan-genome consisted of approximately 14, 000 genes with a core genome of 383 orthologous gene sets. They also highlighted the impressive level of variation in genomic characteristics such as GC content and genome size, ranging from 33% to 51% and 1.8 to 3.3 Mb, respectively (Kant et al., 2011).

The evolutionary and ecological pressures that shape a pan-genome are interactive and multi-dimensional. The processes of gene divergence and gene duplication play a part, whether due to random genetic drift or from positive or purifying selection pressures. These factors cannot explain the impressive level of gene distributions in microbes however - the niche-specific presence of certain genes and the sometimes random scattering of homologous genes across a phylogenetic lineage. Processes that operate on a faster timescale have a greater role in explaining pan-genomes: horizontal gene transfer and gene loss, evolutionary factors that will now be examined in more detail, with examples from *Lactobacillus* species.

3.3 HORIZONTAL GENE TRANSFER AND GENE LOSS

A phylogenetic tree is based on the concept of vertical gene transmission, the passing of genes from parent to daughter cells, be it the action of zygotes following meiosis or the mitotic division of haploid bacterial cells through binary fission. Horizontally transferred genes do not rely on the generational passage of genetic information; they use alternative methods to spread from genome to genome, contradicting the neat description of evolution portrayed by phylogenetic trees. Recent studies have popularised the use of phylogenetic networks in order to capture the horizontal transmission of genes alongside vertical transmission from generation to generation (Huson and Scornavacca, 2011). There are numerous processes that transform the classical notion of a tree into the revolutionary concept of a network, opening up new avenues of research and questioning some of the earlier principles of evolutionary biology.

3.3.1 RECOMBINATION

Muller's ratchet is a concept invoked to explain the evolution of sexual reproduction (Muller, 1964). Sex in eukaryotes prevents the irreversible accumulation of deleterious mutations that would occur in a species that reproduces by purely vertical means. The idea of a mutational ratchet turning in one direction, gradually decreasing the average fitness of an asexually reproducing species is more of a null hypothesis than an evolutionary process that is observed in nature. The closest that biological observations come to Muller's concept of this one-way deterioration of genomes is in endobacteria such as the Mycoplasma-related symbionts that inhabit the cells of fungi. These organisms are vulnerable to genome degeneration because most of the selection pressure acting on their free-living ancestors has been removed by adopting an endosymbiotic lifestyle, but they retain a level of genome plasticity through recombination (Cortez and Weitz, 2014).

Recombination involves the swapping of homologous regions of DNA, catalysed by enzymes such as recombinases. Recombination in eukaryotes ensures that homologous portions of DNA are constantly being reshuffled from generation to

generation, increasing the genetic variation in a population and allowing natural selection to choose from a wider pool of phenotypes (Andersen and Sekelsky, 2010).

In both eukaryotes and prokaryotes, recombination is an important mechanism for DNA repair. Michod *et al* review the adaptive value of sex in microbial pathogens and conclude that recombinational repair of damaged DNA is the main benefit of recombination in pathogens, especially because of the harsh, oxidative environments encountered by pathogens that infect the host cell (Michod et al., 2008). Recombination through processes such as transformation, which involves the taking up of foreign DNA from the environment and its incorporation into the recipient cell, is one way that microbial populations retain their genetic diversity. The classic example of microbial 'sex' however, is displayed by the horizontal transfer of plasmids.

3.3.2 PLASMIDS

The bacterial chromosome contains all the essential genes for a cell to survive and reproduce. Introduce a bacterial strain to a new environment however, and the cells might not have the necessary genes to respond to these altered conditions. Changes to the phenotype of a strain that confer a specific advantage were originally called R-factors (in the case of antibiotic resistance) and similar labels before the term 'plasmid' was coined and later revised to refer to an independently replicating, circular, double-stranded DNA molecule that moved horizontally from cell to cell within microbial populations (Hayes, 2003). The study of these plasmids becomes important when they carry genes that allow bacteria to function in ways for which the chromosome does not code such as the breakdown of a certain sugar or resistance to heavy metals.

Plasmids contain genes that allow them to replicate apart from the chromosome and they can exist anywhere from a single copy to thousands of replicons per cell. Plasmid copy number and plasmid size are usually strongly correlated, with megaplasmids of over 100 kb often being present once per cell, requiring partition proteins to ensure their vertical transfer into both daughter cells, while plasmids as small as 1 kb get transferred in roughly equal proportion in a

probabilistic manner due to sheer number of replicons (Thomas and Summers, 2001).

Plasmids have mechanisms for transferring themselves horizontally from one cell to another. A subset of plasmids have genes involved in the formation of a conjugative "sex" pilus that joins two cells and allows plasmid DNA to be transported across it, endowing a recipient cell with the plasmid-specific properties of the donor. Numerous mobilizable plasmids that do not contain the full complement of genes necessary for conjugation exploit the formation of pili by other plasmids. Some plasmids are thought to be (or have become) completely non-mobilizable, transmitted purely by vertical means, but recent research has downplayed this by providing evidence that 90% of all plasmids in *Staphylococcus aureus* previously thought to be non-mobilizable have mechanisms that assist in HGT by conjugation (Ramsay et al., 2016).

Plasmids are an integral component of genetic diversity and adaptability in *Lactobacillus* species, leading to greater adaptation to different niches and the tendency of *Lactobacillus* pan-genomes to be open. Many of the plasmids in *Lactobacillus* are cryptic, meaning that they have no known function (Wang and Lee, 1997), but numerous plasmids have been identified that increase the functional capacity of their hosts. Ricci *et al* investigated the distribution of plasmids in 22 *L. helveticus* strains isolated from 5 Italian cheeses and found eight plasmid-free strains and multiple plasmids of varying sizes (2.3 to 31 kb) and different homology groups (Ricci et al., 2006). Claesson *et al* described the multireplicon genome architecture of *L. salivarius* UCC118 and showed that a circular megaplasmid expanded the functionality of the strain, coding for a bacteriocin, carbohydrate utilisation genes and a bile salt hydrolase (Claesson et al., 2006). Li *et al* studied 33 strains of *L. salivarius,* showing the ubiquitous presence of the circular megaplasmid ranging in size from 120 kb to 490 kb. Megaplasmids tend to be more stable than smaller plasmids and the study found that phylogenetic comparison of the *repE* gene unique to the megaplasmid followed a similar evolutionary path to the *groEL* gene present on the chromosome, suggesting that this megaplasmid was acquired early in the evolution of *L. salivarius* (Li et al., 2007). Smokvina *et al* describe a plasmid pan-genome of 230 orthologous groups as a subset of the total pan-genome of 4,200 and show that, although a substantial portion of plasmid genes are annotated as

'hypothetical', numerous known adaptive functions are also present (Smokvina et al., 2013).

In genomic studies of extreme environments, plasmids are often the carriers of genes that allow bacterial strains to survive. A study of a Polish copper mine rich in heavy metals showed that plasmids confer resistance to arsenic, cadmium, cobalt, mercury and zinc (Dziewit et al., 2015). Environments with particularly high bacterial densities such as biofilms are hotspots of HGT, including plasmid conjugation. These communities are of particular importance in human health where hospital biofilms promote the transfer of multi-drug resistance to potentially pathogenic organisms (Stalder and Top, 2016).

The existence of plasmids greatly adds to the complexity of evolutionary dynamics in bacteria, increasing gene flow within environmental niches and allowing strains to rapidly adapt to new conditions. There is a stronger ecological force however, one that has been attributed to maintaining microbial population diversity (Olszak et al., 2017) as well as contributing to genetic exchange within and between microbial species (Harrison and Brockhurst, 2017).

3.3.3 BACTERIOPHAGES

Plasmids are usually thought of as beneficial to the host microbe. The common view is to treat them like an adaptation selected at the level of the microbial cell and its genetic lineage, although there are many instances where this is clearly not the case (i.e. cryptic plasmids). This is not true of bacteriophages, viruses composed of either DNA or RNA, encapsulated in protein, that inject their genetic material into the microbial cytoplasm, hijacking cellular machinery for their own replication (McGrath and Van Sinderen, 2007).

Plasmids and bacteriophages exploit microbial cells in very different ways. The plasmid and chromosome often behave symbiotically, both increasing the probability of each other's survival and continued reproduction within the protective structure of the cell. Phages can behave quite violently, multiplying rapidly within their host cell until they are released into the extracellular environment in a chemically-induced burst that signals the death of the cell. This behaviour gives rise to a type of predator-prey cycle where increasing phage numbers lead to a decrease

in the population of the host which in turn provides a lower density of cellular machinery for the phage to manipulate, resulting in a decrease in phage numbers and an increase in host cells, and so on (Cortez and Weitz, 2014).

Phages are divided up into two types, lytic and temperate, depending on whether they have a lytic or a lysogenic cycle. In a lytic cycle, the phage immediately hijacks the transcriptional and translational machinery to make copies of itself, subsequently destroying the host cell and spreading out to infect new hosts. Lysogeny involves the incorporation of the phage genetic material into the chromosome where it lies dormant until unfavourable environmental cues trigger a lytic state (Shao et al., 2017).

In both cycles it is possible for a phage to act as a carrier of bacterial DNA from one host to another, potentially endowing infected cells with new functions. In a lytic state, the formation of bacteriophage capsids can be accompanied by the incorporation of fragments of host DNA into the virion. This happens in a very small number of replicating phages out of the huge number that are released from the cell following lysis, but it is enough to contribute to the horizontal transfer of bacterial genes within a population (Hartl and Jones, 1998), in the phenomenon of transduction.

In a lysogenic state, the bacteriophage genome integrates into the host cell as a prophage and replicates along with the chromosome. In this state, the replication of the prophage is indistinguishable from that of the host genome and both sets of genes are transmitted vertically as the microbial cell divides. When unfavourable environmental conditions trigger excision of the prophage and a return to the lytic state, host DNA can accompany the phage genome due to incorrect excision of the prophage, which then gets packaged into the protective protein capsid along with the phage genome. This process can also leave phage DNA behind in bacterial chromosomes, adding phage genes to the pan-genome of a bacterial species (Harrison and Brockhurst, 2017).

The widespread existence of horizontal gene transfer through plasmid conjugation, phage transduction and recombinational events considerably alters the study of ecology and evolution, especially on a microbial scale where HGT is the dominant mode of adaptation to changing biotic and abiotic conditions. The existence of HGT and the impressive variety of mechanisms involved can reshuffle the list of evolutionary phenomena in order of priority, highlighting concepts that

demand a greater level of explanation if the role of HGT in biology is to be fully appreciated and better understood.

Horizontal gene transfer brings a question sharply into focus, one that has often been overlooked by biologists who study the adaptive evolution of organisms over time. Who or what does the phenotypic expression of a gene benefit? What entity is natural selection acting on to shape each adaptive function in nature? The obvious answer would appear to be the organism with its repertoire of genes, all interacting to achieve a common purpose: survival and reproduction of the members of a species. Selection at the level of organisms has been described as a useful working hypothesis for selection at a lower level, that of the gene, a concept that was first popularised in a 1976 book by Richard Dawkins, *The Selfish Gene*. The book portrays early pre-cellular genes (if they can be called that) as replicators that vary in their longevity, fidelity of replication and rate of replication due to variation in their primary sequence. The banding together of these ancient replicators would have led to evolution of the first cells, very likely in response to the origin of viruses (Forterre, 2006), each replicator/gene having the shared goal of ensuring the phenotypic expression of the genotype successfully interacted with the environment in such a way as to optimise genomic propagation from generation to generation.

The selfish gene theory forces us to pay closer attention to selection acting on horizontally transferred genes, asking whether the strategies of plasmids and maybe even bacteriophages in some sense should primarily be studied as either a cost or a benefit to host bacteria. Richard Dawkins says that plasmid and bacteriophage adaptation should be studied primarily as costing or benefiting the plasmid and bacteriophage genes that code for these adaptations, framing HGT events as adaptive to the genes being transferred, not for the microbial cells receiving the transferred genes.

One particularly large family of genes provides a strong example of the selfish gene theory in action. Transposases, having the apt nickname “jumping genes” among others, possess the ability to cut themselves out of a chromosomal region and paste themselves into other genomic regions, sometimes on extrachromosomal replicons, were they are exported from the cell and transferred to cells of the same and, less commonly, other species (Reznikoff, 2003).

3.3.4 TRANSPOSASES

Transposases, also known as insertion sequences (IS), are the most abundant genes in nature according to a 2010 study that analysed ten million protein-coding genes across bacteria, archaea, eukaryotes and viruses (Aziz et al., 2010). These genes are the pinnacle of selfishness, coding for nothing more than their own horizontal transfer. They multiply within and across genomes and have diversified into a huge variety of groups with many different mechanisms to catalyse transposition.

Transposases do not need homologous recombination in order to insert themselves into a new genomic region; they code for an enzyme that binds to recognised flanking regions, nicking the DNA and forming a complex between the transposase enzymes and the DNA sequence to be transferred to a target site (Reznikoff, 2003). Target DNA sites vary considerably depending on the type of transposase. It was originally thought that transposases show little to no sequence specificity for target regions, but accumulating research is providing evidence against this, showing that some transposases always target a TA dinucleotide sequence (Munoz-Lopez and Garcia-Perez, 2010) while others target longer consensus sequences (Goryshin et al., 1998).

The selfish nature of transposases is an interesting and important area of study in its own right, but the transposase-mediated transfer of additional genes in the form of transposons, pathogenicity islands, antibiotic resistance gene clusters, and other functions is what makes these enzymes key players in HGT. A transposon is simply a gene or group of genes that includes one or more transposases that allow the sequence to be mobilised, often transferring new functions across bacterial strains and species (Darmon and Leach, 2014). Larger clusters of genes known as genomic islands are often flanked by transposases, mobilising entire clusters of genes that would otherwise be confined to the chromosomes of species that shared a common ancestor also possessing the cluster. An excellent example of this in a *Lactobacillus* species is the characterisation of a horizontally transferred operon in *L. curvatus* NRIC0822 of a flagellar operon, conferring motility to the strain. The flagellar motility operon has transposases at both ends and was previously thought to be confined to members of the *L. salivarius* clade before bioinformatic analysis

identified high levels of gene synteny and sequence similarity with the operon in NRIC0822 (Cousin et al., 2015).

There are numerous additional examples in lactobacilli of the role of transposases in the horizontal transfer of novel functions such as genomic islands for cobalamin production in *L. reuteri* (Morita et al., 2008) and tetracycline resistance in *L. sakei* (Devirgiliis et al., 2013). The number of studies that report the mechanisms behind the horizontal acquisition of functions is growing and many of these highlight the role of transposases, as well as plasmids and bacteriophages, in HGT. The contribution of these evolutionary phenomena to the size of bacterial pan-genomes is only beginning to be understood, but there is another process that influences the distribution of genes throughout strains of a species: gene loss, which involves either gene deletion or the inactivation of a gene and its accumulation of mutations over time.

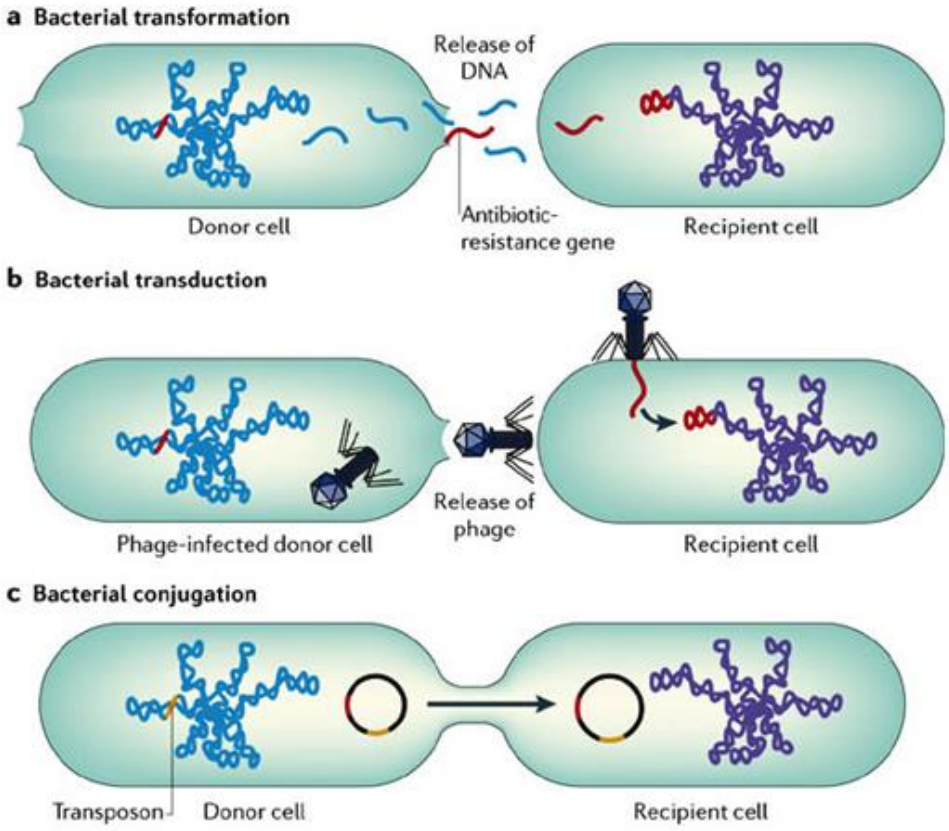


Figure 1.5: The main forms of horizontal gene transfer in bacteria. A) Naked bacterial DNA is released into the environment from a lysed cell and taken up by a recipient cell, where it can be incorporated into the genome. B) Bacteriophage lyse a host cell and infect neighbouring cells by injecting their nucleic acids into the cytoplasm. In a lysogenic state, the phage genome is incorporated into the genome, sometimes carrying bacterial genes with it. C) Plasmid genes code for conjugative structures that act as a bridge between cells in contact, transferring the plasmid from the donor to the recipient cell (Furuya and Lowy, 2006; Springer Nature, License no: 4252530480072).

3.3.5 GENE LOSS

An essential gene, acquiring a mutation (including deletion of the whole gene) that disables the gene's function, will quickly lead to the death of the cell. There is a purifying selection that keeps a species' core genes conserved and fully functional, pruning branches that represent genomes with fatally deleterious

mutations from the phylogenetic tree of dividing cells. No gene loss occurs in this scenario because the loss of any gene means a genome harbouring the lethal mutation will never replicate again, immediately reducing its number of descendants to zero.

An exception to selection acting on core genes to preserve their function is when a gene duplicates. One copy of the gene will remain under purifying selection while the other is free to evolve new functions, although the most common fate of a duplicated gene is to gather mutations until it becomes non-functional, a pseudogene that increasingly loses sequence similarity with its paralogous homolog (Lynch and Conery, 2000).

Gene duplication often leads to a type of gene loss, but the most prevalent form that gene loss takes is when a gene, due to changing environmental conditions, is no longer necessary or even useful for survival, undergoing loss of its function due to genetic drift or active selection pressure. Koskiniemi *et al* provide evidence that selection can be a significant driver of gene loss because unnecessary genes provide a fitness cost to the host. They measured the growth rate of *Salmonella enterica* under multiple conditions involving gene deletions and observed that approximately 25% of deletions led to increased bacterial fitness (Koskiniemi et al., 2012).

Gene loss can also be neutral, resulting from a lack of selection pressure to weed out mutations in genes that no longer confer a fitness benefit to the host. The relative roles of selection and genetic drift in gene loss are still not well understood (Albalat and Canestro, 2016), a varying contribution from both evolutionary forces being likely depending on the environmental context of the gene in question.

Gene loss is often strongly associated with particular niches or reproductive strategies. Endobacteria show considerable gene decay due to decreased selection pressure for nutrients that are easy to access within host cells (Naito and Pawlowska, 2016) while strains of *Lactobacillus casei* isolated from cheese have lost many genes that were no longer needed due to the nutrient-rich environment of the dairy niche (Cai et al., 2009).

The combined effects of HGT and gene loss in shaping the pan-genomes of *Lactobacillus* are mirrored in the gene distributions of many other species. Core genes are involved in a much lower rate of horizontal gene transfer and gene loss while the accessory genome is dominated by both processes, composed of genes from plasmids, bacteriophages, genomic islands and other agents of HGT, which are

usually much more likely to also undergo gene loss (Segerman, 2012). A category of genes that exemplify the dispensable nature of the accessory genome are the strain-specific genes, unique to each genome and quite often un-annotatable, almost as if they reflect an analytical error rather than a biological reality.

3.3.6 STRAIN-SPECIFIC GENES

Describing a gene as strain-specific means nothing without stating the parameters of the dataset to which it belongs (number of strains, species) and the method used to infer all the genes that are not strain-specific. A strain-specific gene can acquire an orthologue if the number of genomes in the dataset is increased, or it can be redefined as part of a group of orthologues if a threshold is lowered (percent identity, for example). Methodological considerations aside, it is very possible for a strain to be the only member of a species possessing a particular gene, having acquired it horizontally from another species in its environment or, less likely, possessing the only remaining gene from an orthologous group that was once common within a species but is now absent from all but one due to gene loss.

Bosi *et al* analysed 64 strains of *Staphylococcus aureus* from a range of niches, host types and antibiotic resistance profiles and found that virulence varied considerably in a strain-dependent manner due to differences in metabolic capabilities (Bosi et al., 2016). In this study, the severity of an *S. aureus* infection was very much contingent on strain-specific genes. The importance of strain-specific properties have been studied in *Lactobacillus* too. Douillard *et al* analysed the genomes of several *L. rhamnosus* and *L. casei* strains and found strain-specific genes with potential probiotic properties (Douillard et al., 2013a). In another study, a genomic comparison of three *L. rhamnosus* strains using probiotic strain GG as a reference revealed strain-specific characteristics with a role in the prevention and possible treatment of *C. difficile* infection (Boonma et al., 2014).

The high numbers of strain-specific genes in most bacteria suggest that horizontal gene transfer and gene loss play a dominant role in the microbial evolution. Gene gain and loss represent the major source of innovation in prokaryotes, occurring at a higher rate than nucleotide substitution (Chang and Duda, 2012). The phylogenetic history of a single gene can contradict that of the majority

of other genes within a genus, but the average statistical tree reflects the phylogenetic history of the information processing gene complexes of the genus as well as other core genes that have been vertically transmitted, free from HGT, acting as a scaffold around which the evolutionary history of all genes can be interpreted (Puigbo et al., 2009). The phylogenetic tree of *Lactobacillus* species and their related genera does not therefore represent the history of all *Lactobacillus* genes, but it is the vertically diversifying evolutionary structure that all HGT events from plasmid conjugation to bacteriophage transduction move within.

3.4 THE PARAPHYLETIC NATURE OF *LACTOBACILLUS*

3.4.1 MONOPHYLY, POLYPHYLY AND PARAPHYLY

Taxonomy is the classification and grouping of organisms according to shared characteristics. Methodological attempts to classify life into groups date as far back as Aristotle, but it was Carl Linnaeus who developed the hierarchical concept of classification and popularised the binomial nomenclature of species that is still in use today. The goal of taxonomy is not to describe the evolutionary relatedness of all living things, but to group them into coherent assemblages of organisms, almost like a catalogue of species for the study of biological disciplines (Grant, 2003).

The advent of phylogenetics signalled the use of molecular information for either the confirmation or contradiction of earlier taxonomic classifications. Phylogeny refers to the evolutionary relationships across organisms, the order of diversification events that occurred within a given lineage that trace all descendants, living and extinct, back to a common ancestor. These relationships are most commonly represented as a tree, each bifurcating branch signifying a speciation event (in the case of actual species formation) or the diversification of lower taxa such as strains (when the phylogeny of a single species is being studied) (Konstantinidis and Tiedje, 2007). Taxonomy and phylogeny can agree on the hierarchical grouping of species, but they can also disagree, in some cases quite significantly. This is because, traditionally, taxonomy focussed on morphological characteristics, which can erroneously make two species appear more closely related than they actually are due to convergent evolution. Phylogenetics, in contrast, utilises

evolutionary signals from molecular data to track the divergence of strains and species from a common ancestor.

A now-famous example of the clashes that can occur between taxonomy and phylogeny is the fact that the hippopotamus is the closest living relative to Cetaceans – whales, dolphins and porpoises. Ursing and Arnason in 1998 confirmed the growing suspicion of this evolutionary relationship by conducting a phylogenetic analysis using the mitochondrial genomes of *Hippopotamus amphibius* and 15 other placental mammals including pigs, horses, cows, sheep and whales (Ursing and Arnason, 1998). These contradictions are dotted throughout the tree of life and have given birth to several terms that describe taxonomic classifications in the light of phylogenetic insight.

A monophyletic clade is a group of organisms classified together under a particular taxonomic name that consists of all the descendants of a common ancestor (Sereno and Lee, 2005). This is the ideal underlying reality of each classified taxon, but human error in deriving imperfect correlations between morphological similarity and evolutionary relatedness leads to complications in the grouping of organisms.

Paraphyletic and polyphyletic clades are two consequences of limitations in taxonomic methodology. Paraphyly involves a group of organisms that share a common ancestor, themselves representing only a subset (usually the majority) of all the descendants of that ancestor. This means that they share the same common ancestor with one or more sub-clades that have a different taxonomic classification. Polyphyly involves a group of organisms that have been classified together according to one or more phenotypic characteristics that do not reflect underlying evolutionary relatedness (Sereno and Lee, 2005). Polyphyletic clades can be scattered in multiple groups across a larger branch of the tree of life and represent both outdated classifications that await systematic modification or groups that have practical use in biological research due to an important phenotypic trait that ties them together.

3.4.2 *LACTOBACILLUS* PHYLOGENY

In *2006*, Canchaya *et al* carried out a comparative genomic analysis of five complete *Lactobacillus* genomes: *L. salivarius*, *L. plantarum*, *L. acidophilus*, *L.*

johnsonii and *L. sakei*. They reported little genome synteny across the five species, which makes sense when they are shown to be scattered across a 16S rRNA gene phylogenetic tree of 111 *Lactobacillus* species. A tree generated from a core protein set of 593 orthologs was largely concordant with a tree generated from whole-genome alignments. The authors concluded that the extreme divergence observed in *Lactobacillus* supported the recognition of sub-generic divisions (Canchaya et al., 2006).

An interesting concept is implicit in the multiple methods of phylogenetic reconstruction used by Canchaya *et al* (Canchaya et al., 2006): there is a phylogenetic tree and there is an underlying phylogeny, the former being an output from a particular method of organismal comparison that takes evolutionary patterns into account, the latter representing the actual evolutionary history of the genomes being analysed. A phylogenetic tree is an estimation of the true underlying series of speciation events of a group of organisms and different metrics as well as different regions of the genome can and do disagree, both because of variations in algorithmic and biological assumptions, and variations in phylogenetic signal in different DNA sequences.

Claesson *et al* noted in 2007 that numerous species of *Lactobacillus* have been reclassified to other genera and the taxonomy of the genus at the time was generally unsatisfactory (Claesson et al., 2007). Makarova & Koonin in 2007 stated that classification of *Lactobacillales* remained an unresolved issue because the phenotypic scheme for taxonomic assignment was based on fermentation profiles and disagreed with rRNA-based phylogeny. A phylogenetic tree constructed from four subunits of the DNA-dependent RNA polymerase showed *Pediococcus*, *Leuconostoc* and *Oenococcus* branching from within *Lactobacillus* (Makarova and Koonin, 2007).

A 2008 study by Claesson *et al* used several whole-genome and single-gene phylogenetic analyses in an attempt to sub-divide lactobacilli into coherent sub-generic groups. They found significant incongruencies among phylogenetic analyses and hypothesised that these are due to differences in evolutionary rates, hidden paralogies (mistaken orthology) and HGT. They showed that the GroEL gene is a more robust phylogenetic marker than the 16S rRNA gene for single-gene phylogeny of lactobacilli and, despite contradictions in the clustering of sub-clades, four sub-generic groups showed considerable robustness. Interestingly, they found that these

major groupings were more clearly defined by gene absence than the presence of gained genes, highlighting the trend toward genome reduction found in lactobacilli due to their adaptation to high-nutrient environments (Claesson et al., 2008).

Kant *et al* constructed a phylogenetic tree based on a core genome of 383 orthologues from 20 complete *Lactobacillus* genomes, defining separate groups based on group-specific core genes. The aim of this study was to provide a platform for present and future analyses involving *Lactobacillus* genomes, acting as a kind of template for the addition of new species (Kant et al., 2011).

It was conclusions like those of the Claesson *et al*, Makarova & Koonin and Kant *et al* studies that led Salvetti *et al* to update the phylogenetic tree of *Lactobacillus* based on the 16S rRNA gene, dividing the genus up into sub-clades consisting of 15 groups of three or more species, four pairs and 10 single lines of descent. They also noted that the genus *Pediococcus* branches from within *Lactobacillus*, confirming previous evidence that the lactobacilli are a paraphyletic genus (Salvetti et al., 2012). This study was based on the 16S rRNA gene and it is interesting to consider how similar their tree topology would be if different marker genes were used.

Lukjancenko *et al* showed that *Leuconostoc* branches from within *Lactobacillus* when species are clustered based on variable gene content (Lukjancenko et al., 2012). Variable gene content correlates well with core- and marker-gene phylogeny at a species level because the accumulation of gene loss and HGT events occurs over time as core-gene sequences diverge.

The number of recognised *Lactobacillus* genomes is continuously growing, rising with the publication of each new study as research expands in the areas of food fermentation and probiotics. From around 80 species in 2006 (Canchaya et al., 2006) to 152 in 2012 (Salvetti et al., 2013), *Lactobacillus* is a genus with a regular stream of new members. Holzapfel *et al* emphasised the explosion of new species discovered over a 15-year period up to 2014, highlighting the tendency for early phenotypic classifications to be transferred to newly created genera, including *Atopobium*, *Carnobacterium*, *Eggerthia*, *Fructobacillus* and *Weissella*. They concluded that the phylogenetic diversity of *Lactobacillus* warrants genotypic subdivision of the genus, a conclusion that has been put forward by previous studies (Holzapfel and Wood, 2014).

In 2015, Goldstein *et al* reported the number of recognised *Lactobacillus* species as 170 and stressed that they cannot be differentiated easily by phenotypic means. They noted that the antimicrobial susceptibility of *Lactobacillus* species was poorly defined in large part because of their taxonomic complexity (Goldstein et al., 2015). It is this taxonomic complexity and the impressive levels of phylogenetic and functional diversity that make *Lactobacillus* such an interesting genus for researchers, even when their importance in the food industry and in human health is not the focus of study.

The extensive horizontal transfer of genes and gene loss events in *Lactobacillus* reflects the range of niches that they occupy and the varying selection pressure that must act on their genes as they adapt to new and changing environments. Variation in gene presence across a taxon is an important phenomenon to understand, elucidating the different processes involved in the evolution of bacterial species. The evolution of core genes is also informative, the analysis of sequence divergence revealing information on the role of varying selection pressure throughout genomic regions as well as along the length of each gene, constraining certain amino acid residues while allowing others to mutate and become fixed within a bacterial population, both through genetic drift and positive selection pressure. The evolutionary rate and associated factors are at the heart of these studies because an explanation of evolutionary rate variation across lineages and within genes provides insight into the evolutionary forces that have acted on, and are still acting on, the genomes of organisms.

3.5 EVOLUTIONARY RATE

3.5.1 MUTATION RATE

Mutations are the nucleotide base changes that accumulate in a DNA sequence over time. They can happen as point mutations, leading to single base changes or single base insertion or deletion events, or they can involve the inversion or translocation of larger sequence regions. Mutations in DNA occur because the copying fidelity of cellular replication machinery is less than perfect. Endogenous factors such as reactive oxygen species and exogenous factors like UV light also

cause mutations through various processes. Enzymes involved in DNA repair reduce the number of these mutations, often by using homologous nucleotide regions that still code for the correct sequence. DNA repair mechanisms are also error prone, which means that no actively reproducing organism is ever really mutation-free (Bertram, 2000).

A mutation rate of zero would mean that all descendant sequences are identical to an ancestral sequence, leading to the absence of evolution and nothing on which natural selection can act. When asked to define evolution in a sentence, Richard Dawkins stated that "Life results from the non-random survival of randomly varying replicators." Genetic variation is necessary for evolution, and mutation is the process that restocks the sequence variation lost through genetic drift, purifying selection and the fixation of gene variants within a population.

Mutation rate is not the same as fixation rate. A mutation occurs in the larger context of a population of organisms and the gene variant produced by the mutation initially has a frequency of one. The gene variant is said to be "fixed" when it is found in every member of the population. A neutral mutation can drift to fixation through random oscillations in frequency while deleterious mutations can become fixed in a relatively small population when its negative effect on fitness is small. An advantageous mutation will become fixed at a rate determined by the increase in fitness of the genomes it occupies, taking effective population size into account.

The mutation rate is far from uniform. Sniegowski *et al* noted that selection can adjust the mutation rate by acting on sequence variation in genes responsible for DNA replication and repair. They hypothesise that, since most mutations are either neutral or deleterious, the mutation rate of a species is as low as the physiological cost of increased fidelity will allow, concluding that selection for higher mutation rates is likely only in special cases (Sniegowski et al., 2000). Such a case surely exists in the bacterial pathogen, *Helicobacter pylori*, which was shown to have a mutation rate over 10 times faster during the acute phase versus the chronic phase of infection in humans and rhesus macaques. The elevated mutation rate of *H. pylori* during acute infection is orders of magnitude faster than any other studied bacterium and likely facilitates rapid adaptation to the host environment (Linz et al., 2014). In contrast to the "cost of fidelity" hypothesis put forward by Sniegowski *et al*, Lynch supports the hypothesis that the lower limit imposed on the mutation rate is explained by genetic drift (Lynch, 2010).

Wielgoss *et al* conducted an evolutionary experiment over 40,000 generations of *E. coli* in order to quantify the spontaneous mutation rate in this species. They sequenced 19 *E. coli* genomes at the end of the experiment and directly inferred the point mutation rate based on accumulated substitutions, calculating a rate of 8.9 $x10^{-11}$ per base-pair per generation and recording a significant bias toward increased AT content (Lynch, 2010). The measured substitutions were limited to a particular subset of mutations - those that occur within protein-coding genes but code for the same amino acid due to the redundancy of the genetic code.

3.5.2 SYNONYMOUS AND NON-SYNONYMOUS MUTATIONS

Each amino acid of a gene is coded for by a triplet of nucleotides, but not every triplet codes for a different amino acid. There are $4^3 = 64$ possible codon triplets and only 20 amino acids, and every possible triplet codes for either an amino acid or a stop codon. The genetic code leads to a redundancy where several different codons can be translated into the same amino acid, usually those sharing the first two nucleotide bases (Watson, 1970).

A mutation in a gene sequence that leaves the translated amino acid sequence unaltered is called a synonymous mutation while one that leads to an amino acid change is called a non-synonymous mutation. It is a common assumption that synonymous mutations are hidden from natural selection because they leave the protein sequence, and therefore the phenotype, unaltered. For this reason, the rate of synonymous mutation is assumed to reflect the mutation rate because both operate in the absence of selection (Zhang and Yang, 2015). However, it has been shown that synonymous mutations do have an influence on the 'genome phenotype', the tendency for changes in nucleotide sequences to affect transcription and translation accuracy as well as the rate of protein mis-folding and a range of other processes (Forsdyke, 2002).

Non-synonymous mutations represent the mutation rate under selective pressure, whether it is a conserved structural protein domain under purifying selection or an active protein site displaying considerable amino acid variation across species due to strong positive selection. For a given gene, the number of potential

synonymous and non-synonymous sites can vary, which can lead to false conclusions about the evolutionary rate or the type of selection acting on the sequence. Normalised values are more commonly used where dN = number of non-synonymous substitutions per non-synonymous site and dS = number of synonymous substitutions per synonymous site. The value for dS is used as a relative mutation rate for the gene and the ratio, dN/dS, reflects the evolutionary rate of proteins normalised for variation in mutation rate under a neutral model and can be used as a measure of the strength and type of selection pressure acting on a gene (Zhang and Yang, 2015).

A dN/dS value of approximately one suggests that a gene is under, on average, neutral selection pressure since the proportion of substitutions subject to selection and the proportion hidden from it both accumulate at the same rate. A value of less than one suggests that purifying selection is constraining amino acid substitutions while synonymous mutations occur unchecked by selection. A value of greater than one suggests that non-synonymous mutations are positively selected for relative to neutral synonymous mutations, which follow a statistical fixation or elimination implied by genetic drift (Zhang and Yang, 2015).

3.5.3 SELECTION PRESSURE AND EVOLUTIONARY RATE

The evolutionary rate of a gene is a measure of how fast its sequence evolves. Excluding pseudogenes, which are practically selectively neutral, the divergence of two nucleotide sequences will occur much more quickly than their corresponding amino acid translation. Evolutionary rate therefore depends very much on whether it is being measured at the nucleotide or protein level.

A protein sequence evolves at a uniform rate neither along its length nor over time. In a given environment, selection acts on each individual codon of a gene, constraining amino acids that are essential for protein function and allowing other residues to vary once overall protein structure is not compromised (Yang, 1996). If environmental conditions change, selection pressure on a gene may change and each amino acid will potentially be affected by an altered pressure. It is difficult to predict the effect that a specific environmental change will have on evolutionary rate, but there is evidence to suggest that changing environments, particularly unpredictable

ones, favour adaptations that increase the mutation rate, leading to positive selection for a subset of non-synonymous mutations, which increase the average evolutionary rate of a population (Denamur and Matic, 2006).

The average selection pressure acting on most genes may well be purifying, preventing most amino acid changes from persisting in a population, but bacterial genes involved in overcoming host defences have been shown to be under positive selection pressure due to the evolutionary arms race of adaptation and counter-adaptation between host and pathogen (Jordan et al., 2002). This has been shown in pathogenic strains of *E. coli* where cell-surface proteins are under positive selection because of their interaction with the changing environment of the host (Petersen et al., 2007).

Functional importance of a protein was once thought to be the main factor affecting the evolutionary rate of the gene that codes for it - the more important the protein, the slower the rate of evolution. It has been found that expression level is the major determinant of evolutionary rate, with functional importance playing only a minor role (Zhang and Yang, 2015). Numerous other factors have been shown to correlate with evolutionary rate and also with each other, making the phrase 'correlation versus causation' very much central to the interpretation of results.

The complexity of analyses involving evolutionary rate is further increased by the array of methods that are currently used to measure it. The first step involves the multiple alignment of homologous genes (either nucleotides or amino acids) carried out using software such as Muscle (Edgar, 2010) and CLUSTALW (Thompson et al., 1994), which attempt to introduce gap positions in order to preserve positional homology across sequences. Amino acid substitution matrices are used to assign similarity scores to sequence alignments based on the agreement of physico-chemical properties between homologous positions (Henikoff and Henikoff, 1992) while values of dN and dS are calculated from aligned homologous codons. Maximum parsimony, maximum likelihood and Bayesian methods are also used to calculate evolutionary rate (Bevan et al., 2005).

The *Lactobacillus* genus is phylogenetically and functionally diverse, making the study of evolutionary rate across its genes both intriguing and daunting. Makarova & Koonin used a molecular-clock test on a phylogenetic tree generated from ribosomal proteins to reveal a high heterogeneity of evolutionary rates among *Lactobacillales* (Makarova and Koonin, 2007). This finding is not so surprising

given that *Lactobacillus* species are scattered over such a variety of niches and wide array of environmental conditions. Attaching absolute rates to evolutionary events is difficult, often relying on simplistic assumptions in the absence of fossil and other temporal evidence. An alternative approach is to use a relative measure of evolutionary rate such as the dN/dS ratio that normalises for variation in mutation rate across genes, but cannot provide estimates of the occurrence of speciation events in time (Zhang and Yang, 2015).

New *Lactobacillus* species are being announced every year, accompanied by the increasing rate of published studies on lactobacilli. Comparative genomic and phylogenomic studies of lactobacilli are also becoming more numerous, fuelled by the rapidly falling cost of sequencing and the expansion of bioinformatic tools designed specifically for these types of analyses. Future studies will continue to reveal the impressive level of functional variation characteristic of the *Lactobacillus* genus, providing greater insight into its evolutionary and ecological dynamics. The potential consequences include advances in human health through probiotics, increased efficiency of food preservation and advances in the industrial use of lactobacilli.

4 BIBLIOGRAPHY

ALBALAT, R. & CANESTRO, C. 2016. Evolution by gene loss. *Nat Rev Genet,* 17**,** 379-91.

ALBERTS, B. 2015. *Molecular biology of the cell*.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ANDERSEN, S. L. & SEKELSKY, J. 2010. Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *Bioessays,* 32**,** 1058-66.

AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med,* 79**,** 137-58.

AZIZ, R. K., BREITBART, M. & EDWARDS, R. A. 2010. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res,* 38**,** 4207-17.

BAIROCH, A. & APWEILER, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* 28**,** 45-8.

BAKER, M. 2012. De novo genome assembly: what every biologist should know. *Nat Meth,* 9**,** 333-337.

BALZER, S., MALDE, K. & JONASSEN, I. 2011. Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics,* 27**,** i304-9.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

BERTRAM, J. S. 2000. The molecular biology of cancer. *Mol Aspects Med,* 21**,** 167-223.

BEVAN, R. B., LANG, B. F. & BRYANT, D. 2005. Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Syst Biol,* 54**,** 900-15.

BOETZER, M. & PIROVANO, W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol,* 13**,** R56.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

BOOLE, G. 1847. *The Mathematical Analysis of Logic, Being an Essay towards a Calculus of Deductive Reasoning,* London, England, Macmillan, Barclay, & Macmillan.

BOONMA, P., SPINLER, J. K., QIN, X., JITTAPRASATSIN, C., MUZNY, D. M., DODDAPANENI, H., GIBBS, R., PETROSINO, J., TUMWASORN, S. & VERSALOVIC, J. 2014. Draft genome sequences and description of Lactobacillus rhamnosus strains L31, L34, and L35. *Stand Genomic Sci,* 9**,** 744-54.

BOSI, E., MONK, J. M., AZIZ, R. K., FONDI, M., NIZET, V. & PALSSON, B. O. 2016. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc Natl Acad Sci U S A,* 113**,** E3801-9.

BROADBENT, J. R., NEENO-ECKWALL, E. C., STAHL, B., TANDEE, K., CAI, H., MOROVIC, W., HORVATH, P., HEIDENREICH, J., PERNA, N. T., BARRANGOU, R. & STEELE, J. L. 2012. Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. *BMC Genomics,* 13**,** 533.

BURKHARDT, R. W., JR. 2013. Lamarck, evolution, and the inheritance of acquired characters. *Genetics,* 194**,** 793-805.

CAI, H., THOMPSON, R., BUDINICH, M. F., BROADBENT, J. R. & STEELE, J. L. 2009. Genome sequence and comparative genome analysis of Lactobacillus casei: insights into their niche-associated evolution. *Genome Biol Evol,* 1**,** 239-57.

CANCHAYA, C., CLAESSON, M. J., FITZGERALD, G. F., VAN SINDEREN, D. & O'TOOLE, P. W. 2006. Diversity of the genus Lactobacillus revealed by comparative genomics of five species. *Microbiology,* 152**,** 3185-96.

CHANG, D. & DUDA, J. T. F. 2012. Extensive and Continuous Duplication Facilitates Rapid Evolution and Diversification of Gene Families. *Molecular Biology and Evolution,* 29**,** 2019-2029.

CLAESSON, M. J., LI, Y., LEAHY, S., CANCHAYA, C., VAN PIJKEREN, J. P., CERDENO-TARRAGA, A. M., PARKHILL, J., FLYNN, S., O'SULLIVAN, G. C., COLLINS, J. K., HIGGINS, D., SHANAHAN, F., FITZGERALD, G. F., VAN SINDEREN, D. & O'TOOLE, P. W. 2006. Multireplicon genome architecture of Lactobacillus salivarius. *Proc Natl Acad Sci U S A,* 103**,** 6718-23.

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE, P. W. 2007. The genus Lactobacillus--a genomic basis for understanding its diversity. *FEMS Microbiol Lett,* 269**,** 22-8.

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE, P. W. 2008. Lactobacillus phylogenomics--towards a reclassification of the genus. *Int J Syst Evol Microbiol,* 58**,** 2945-54.

COMPEAU, P. E. C., PEVZNER, P. A. & TESLER, G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotech,* 29**,** 987-991.

CORTEZ, M. H. & WEITZ, J. S. 2014. Coevolution can reverse predator-prey cycles. *Proc Natl Acad Sci U S A,* 111**,** 7486-91.

COUSIN, F. J., LYNCH, S. M., HARRIS, H. M., MCCANN, A., LYNCH, D. B., NEVILLE, B. A., IRISAWA, T., OKADA, S., ENDO, A. & O'TOOLE, P. W. 2015. Detection and genomic characterization of motility in Lactobacillus curvatus: confirmation of motility in a species outside the Lactobacillus salivarius clade. *Appl Environ Microbiol,* 81**,** 1297-1308.

CRISTIANINI, N. & HAHN, M. W. 2007. *Introduction to Computational Genomics: A Case Studies Approach*, Cambridge University Press.

DARMON, E. & LEACH, D. R. 2014. Bacterial genome instability. *Microbiol Mol Biol Rev,* 78**,** 1-39.

DASGUPTA, S. 2016. *Computer Science: A Very Short Introduction*, Oxford University Press.

DEL FABBRO, C., SCALABRIN, S., MORGANTE, M. & GIORGI, F. M. 2013. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One,* 8**,** e85024.

DENAMUR, E. & MATIC, I. 2006. Evolution of mutation rates in bacteria. *Mol Microbiol,* 60**,** 820-7.

DESIERE, F., PRIDMORE, R. D. & BRUSSOW, H. 2000. Comparative genomics of the late gene cluster from Lactobacillus phages. *Virology,* 275**,** 294-305.

DEVIRGILIIS, C., ZINNO, P. & PEROZZI, G. 2013. Update on antibiotic resistance in foodborne Lactobacillus and Lactococcus species. *Front Microbiol,* 4**,** 301.

DOBZHANSKY, T. 1973. Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher,* 35**,** 125-129.

DONATI, C., HILLER, N. L., TETTELIN, H., MUZZI, A., CROUCHER, N. J., ANGIUOLI, S. V., OGGIONI, M., DUNNING HOTOPP, J. C., HU, F. Z., RILEY, D. R., COVACCI, A., MITCHELL, T. J., BENTLEY, S. D., KILIAN, M., EHRLICH, G. D., RAPPUOLI, R., MOXON, E. R. & MASIGNANI, V. 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol,* 11**,** R107.

DOUILLARD, F. P., RIBBERA, A., JARVINEN, H. M., KANT, R., PIETILA, T. E., RANDAZZO, C., PAULIN, L., LAINE, P. K., CAGGIA, C., VON OSSOWSKI, I., REUNANEN, J., SATOKARI, R., SALMINEN, S., PALVA, A. & DE VOS, W. M. 2013a. Comparative genomic and functional analysis of Lactobacillus casei and Lactobacillus rhamnosus strains marketed as probiotics. *Appl Environ Microbiol,* 79**,** 1923-33.

DOUILLARD, F. P., RIBBERA, A., KANT, R., PIETILA, T. E., JARVINEN, H. M., MESSING, M., RANDAZZO, C. L., PAULIN, L., LAINE, P., RITARI, J., CAGGIA, C., LAHTEINEN, T., BROUNS, S. J., SATOKARI, R., VON OSSOWSKI, I., REUNANEN, J., PALVA, A. & DE VOS, W. M. 2013b. Comparative genomic and functional analysis of 100 Lactobacillus rhamnosus strains and their comparison with strain GG. *PLoS Genet,* 9**,** e1003683.

DWORKIN, M. 2012. Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. *FEMS Microbiol Rev,* 36**,** 364-79.

DZIEWIT, L., PYZIK, A., SZUPLEWSKA, M., MATLAKOWSKA, R., MIELNICKI, S., WIBBERG, D., SCHLUTER, A., PUHLER, A. & BARTOSIK, D. 2015. Diversity and role of plasmids in adaptation of bacteria inhabiting the Lubin copper mine in Poland, an environment rich in heavy metals. *Front Microbiol,* 6**,** 152.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-1.

EKBLOM, R. & WOLF, J. B. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl,* 7**,** 1026-42.

FIERS, W., CONTRERAS, R., DUERINCK, F., HAEGEMAN, G., ISERENTANT, D., MERREGAERT, J., MIN JOU, W., MOLEMANS, F., RAEYMAEKERS, A., VAN DEN BERGHE, A., VOLCKAERT, G. & YSEBAERT, M. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature,* 260**,** 500-7.

FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J. & BATEMAN, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res,* 44**,** D279-85.

FORSDYKE, D. R. 2002. Selective pressures that decrease synonymous mutations in Plasmodium falciparum. *Trends Parasitol,* 18**,** 411-7.

FORTERRE, P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res,* 117**,** 5-16.

FRESE, S. A., BENSON, A. K., TANNOCK, G. W., LOACH, D. M., KIM, J., ZHANG, M., OH, P. L., HENG, N. C., PATIL, P. B., JUGE, N., MACKENZIE, D. A., PEARSON, B. M., LAPIDUS, A., DALIN, E., TICE, H., GOLTSMAN, E., LAND, M., HAUSER, L., IVANOVA, N., KYRPIDES, N. C. & WALTER, J. 2011. The evolution of host specialization in the vertebrate gut symbiont Lactobacillus reuteri. *PLoS Genet,* 7**,** e1001314.

FURUYA, E. Y. & LOWY, F. D. 2006. Antimicrobial-resistant bacteria in the community setting. *Nat Rev Micro,* 4**,** 36-45.

GOLDSTEIN, E. J., TYRRELL, K. L. & CITRON, D. M. 2015. Lactobacillus species: taxonomic complexity and controversial susceptibilities. *Clin Infect Dis,* 60 Suppl 2**,** S98-107.

GORYSHIN, I. Y., MILLER, J. A., KIL, Y. V., LANZOV, V. A. & REZNIKOFF, W. S. 1998. Tn5/IS50 target recognition. *Proc Natl Acad Sci U S A,* 95**,** 10716-21.

GRANT, V. 2003. Incongruence between cladistic and taxonomic systems. *American Journal of Botany,* 90**,** 1263-1270.

GRATH, S. M. & SINDEREN, D. V. 2007. *Bacteriophage: Genetics and Molecular Biology*, Caister Academic Press.

HAAS, L. F. 1994. Charles Babbage (1792-1871). *J Neurol Neurosurg Psychiatry,* 57**,** 1025.

HAFT, D. H., SELENGUT, J. D. & WHITE, O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res,* 31**,** 371-3.

HARRISON, E. & BROCKHURST, M. A. 2017. Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn't Kill You Makes You Stronger. *Bioessays*.

HARTL, D. L. & JONES, E. W. 1998. *Genetics: Principles and Analysis*, Jones and Bartlett Publishers.

HAYES, F. 2003. The function and organization of plasmids. *Methods Mol Biol,* 235**,** 1-17.

HAYES, W. S. & BORODOVSKY, M. 1998. How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res,* 8**,** 1154-71.

HEATHER, J. M. & CHAIN, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

HENIKOFF, S. & HENIKOFF, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* 89**,** 10915-9.

HIRSCH, C. N., FOERSTER, J. M., JOHNSON, J. M., SEKHON, R. S., MUTTONI, G., VAILLANCOURT, B., PENAGARICANO, F., LINDQUIST, E., PEDRAZA, M. A., BARRY, K., DE LEON, N., KAEPPLER, S. M. & BUELL, C. R. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell,* 26**,** 121-35.

HOLZAPFEL, W. H. & WOOD, B. J. B. 2014. Introduction to the LAB. *Lactic Acid Bacteria.* John Wiley & Sons, Ltd.

HUSON, D. H. & SCORNAVACCA, C. 2011. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol,* 3**,** 23-35.

JACOB, E., HOROVITZ, A. & UNGER, R. 2007. Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study. *Bioinformatics,* 23**,** i240-8.

JORDAN, I. K., ROGOZIN, I. B., WOLF, Y. I. & KOONIN, E. V. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res,* 12**,** 962-8.

KANT, R., BLOM, J., PALVA, A., SIEZEN, R. J. & DE VOS, W. M. 2011. Comparative genomics of Lactobacillus. *Microb Biotechnol,* 4**,** 323-32.

KANT, R., RINTAHAKA, J., YU, X., SIGVART-MATTILA, P., PAULIN, L., MECKLIN, J. P., SAARELA, M., PALVA, A. & VON OSSOWSKI, I. 2014. A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in Lactobacillus rhamnosus. *PLoS One,* 9**,** e102762.

KONSTANTINIDIS, K. T. & TIEDJE, J. M. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol,* 10**,** 504-9.

KOSKINIEMI, S., SUN, S., BERG, O. G. & ANDERSSON, D. I. 2012. Selection-driven gene loss in bacteria. *PLoS Genet,* 8**,** e1002787.

LAGESEN, K., HALLIN, P., RODLAND, E. A., STAERFELDT, H. H., ROGNES, T. & USSERY, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res,* 35**,** 3100-8.

LAND, M., HAUSER, L., JUN, S. R., NOOKAEW, I., LEUZE, M. R., AHN, T. H., KARPINETS, T., LUND, O., KORA, G., WASSENAAR, T., POUDEL, S. & USSERY, D. W. 2015. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics,* 15**,** 141-61.

LANE, N. 2015. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci,* 370.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LI, R., LI, Y., ZHENG, H., LUO, R., ZHU, H., LI, Q., QIAN, W., REN, Y., TIAN, G., LI, J., ZHOU, G., ZHU, X., WU, H., QIN, J., JIN, X., LI, D., CAO, H., HU, X., BLANCHE, H., CANN, H., ZHANG, X., LI, S., BOLUND, L., KRISTIANSEN, K., YANG, H., WANG, J. & WANG, J. 2010. Building the sequence map of the human pan-genome. *Nat Biotechnol,* 28**,** 57-63.

LI, Y., CANCHAYA, C., FANG, F., RAFTIS, E., RYAN, K. A., VAN PIJKEREN, J. P., VAN SINDEREN, D. & O'TOOLE, P. W. 2007. Distribution of megaplasmids in Lactobacillus salivarius and other lactobacilli. *J Bacteriol,* 189**,** 6128-39.

LINZ, B., WINDSOR, H. M., MCGRAW, J. J., HANSEN, L. M., GAJEWSKI, J. P., TOMSHO, L. P., HAKE, C. M., SOLNICK, J. V., SCHUSTER, S. C. & MARSHALL, B. J. 2014. A mutation burst during the acute phase of Helicobacter pylori infection in humans and rhesus macaques. *Nat Commun,* 5**,** 4165.

LOWE, T. M. & EDDY, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res,* 25**,** 955-64.

LU, H., GIORDANO, F. & NING, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics,* 14**,** 265-279.

LUKASHIN, A. V. & BORODOVSKY, M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res,* 26**,** 1107-15.

LUKJANCENKO, O., USSERY, D. W. & WASSENAAR, T. M. 2012. Comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera. *Microb Ecol,* 63**,** 651-73.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y., TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, D. W., YIU, S. M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T. W. & WANG, J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience,* 1**,** 18.

LYNCH, M. 2010. Evolution of the mutation rate. *Trends Genet,* 26**,** 345-52.

LYNCH, M. & CONERY, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science,* 290**,** 1151-5.

MADIGAN, M. T. 2012. *Brock biology of microorganisms*.

MAKAROVA, K., SLESAREV, A., WOLF, Y., SOROKIN, A., MIRKIN, B., KOONIN, E., PAVLOV, A., PAVLOVA, N., KARAMYCHEV, V., POLOUCHINE, N., SHAKHOVA, V., GRIGORIEV, I., LOU, Y., ROHKSAR, D., LUCAS, S., HUANG, K., GOODSTEIN, D. M., HAWKINS, T., PLENGVIDHYA, V., WELKER, D., HUGHES, J., GOH, Y., BENSON, A., BALDWIN, K., LEE, J. H., DIAZ-MUNIZ, I., DOSTI, B., SMEIANOV, V., WECHTER, W., BARABOTE, R., LORCA, G., ALTERMANN, E., BARRANGOU, R., GANESAN, B., XIE, Y., RAWSTHORNE, H., TAMIR, D., PARKER, C., BREIDT, F., BROADBENT, J., HUTKINS, R., O'SULLIVAN, D., STEELE, J., UNLU, G., SAIER, M., KLAENHAMMER, T., RICHARDSON, P., KOZYAVKIN, S., WEIMER, B. & MILLS, D. 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A,* 103**,** 15611-6.

MAKAROVA, K. S. & KOONIN, E. V. 2007. Evolutionary genomics of lactic acid bacteria. *J Bacteriol,* 189**,** 1199-208.

MARTINO, M. E., BAYJANOV, J. R., CAFFREY, B. E., WELS, M., JONCOUR, P., HUGHES, S., GILLET, B., KLEEREBEZEM, M., VAN HIJUM, S. A. & LEULIER, F. 2016. Nomadic lifestyle of Lactobacillus plantarum revealed by comparative genomics of 54 strains isolated from different habitats. *Environ Microbiol,* 18**,** 4974-4989.

MENDEL, G. 1866. Experiments on plant hybridisation. *Verhandlungen des naturforschenden Vereins Brünn*.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

MICHOD, R. E., BERNSTEIN, H. & NEDELCU, A. M. 2008. Adaptive value of sex in microbial pathogens. *Infect Genet Evol,* 8**,** 267-85.

MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95**,** 315-27.

MIN JOU, W., HAEGEMAN, G., YSEBAERT, M. & FIERS, W. 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature,* 237**,** 82-8.

MOORTHIE, S., MATTOCKS, C. J. & WRIGHT, C. F. 2011. Review of massively parallel DNA sequencing technologies. *Hugo J,* 5**,** 1-12.

MORAES, F. & GOES, A. 2016. A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochem Mol Biol Educ,* 44**,** 215-23.

MORITA, H., TOH, H., FUKUDA, S., HORIKAWA, H., OSHIMA, K., SUZUKI, T., MURAKAMI, M., HISAMATSU, S., KATO, Y., TAKIZAWA, T., FUKUOKA, H., YOSHIMURA, T., ITOH, K., O'SULLIVAN, D. J., MCKAY, L. L., OHNO, H., KIKUCHI, J., MASAOKA, T. & HATTORI, M. 2008. Comparative genome analysis of Lactobacillus reuteri and Lactobacillus fermentum reveal a genomic island for reuterin and cobalamin production. *DNA Res,* 15**,** 151-61.

MULLER, H. J. 1964. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutat Res,* 106**,** 2-9.

MUNOZ-LOPEZ, M. & GARCIA-PEREZ, J. L. 2010. DNA transposons: nature and applications in genomics. *Curr Genomics,* 11**,** 115-28.

NAITO, M. & PAWLOWSKA, T. E. 2016. Defying Muller's Ratchet: Ancient Heritable Endobacteria Escape Extinction through Retention of Recombination and Genome Plasticity. *MBio,* 7.

NOGUCHI, H., PARK, J. & TAKAGI, T. 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res,* 34**,** 5623-30.

NYQUIST, O. L., MCLEOD, A., BREDE, D. A., SNIPEN, L., AAKRA, A. & NES, I. F. 2011. Comparative genomics of Lactobacillus sakei with emphasis on strains from meat. *Mol Genet Genomics,* 285**,** 297-311.

NYREN, P. 2015. The History of Pyrosequencing((R)). *Methods Mol Biol,* 1315**,** 3-15.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res,* 27**,** 29-34.

OJALA, T., KANKAINEN, M., CASTRO, J., CERCA, N., EDELMAN, S., WESTERLUND-WIKSTROM, B., PAULIN, L., HOLM, L. & AUVINEN, P. 2014. Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics,* 15**,** 1070.

OLSZAK, T., LATKA, A., ROSZNIOWSKI, B., VALVANO, M. A. & DRULIS-KAWA, Z. 2017. Phage life cycles behind bacterial biodiversity. *Curr Med Chem*.

PETERSEN, L., BOLLBACK, J. P., DIMMIC, M., HUBISZ, M. & NIELSEN, R. 2007. Genes under positive selection in Escherichia coli. *Genome Res,* 17**,** 1336-43.

PROUX, C., VAN SINDEREN, D., SUAREZ, J., GARCIA, P., LADERO, V., FITZGERALD, G. F., DESIERE, F. & BRUSSOW, H. 2002. The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J Bacteriol,* 184**,** 6026-36.

PUIGBO, P., WOLF, Y. I. & KOONIN, E. V. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol,* 8**,** 59.

QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G., MIKHAIL, A., OUEDRAOGO, N., AFROUGH, B., BAH, A., BAUM, J. H., BECKER-ZIAJA, B., BOETTCHER, J. P., CABEZA-CABRERIZO, M., CAMINO-SANCHEZ, A., CARTER, L. L., DOERRBECKER, J., ENKIRCH, T., DORIVAL, I. G. G., HETZELT, N., HINZMANN, J., HOLM, T., KAFETZOPOULOU, L. E., KOROPOGUI, M., KOSGEY, A., KUISMA, E., LOGUE, C. H., MAZZARELLI, A., MEISEL, S., MERTENS, M., MICHEL, J., NGABO, D., NITZSCHE, K., PALLASH, E., PATRONO, L. V., PORTMANN, J., REPITS, J. G., RICKETT, N. Y., SACHSE, A., SINGETHAN, K., VITORIANO, I., YEMANABERHAN, R. L., ZEKENG, E. G., TRINA, R., BELLO, A., SALL, A. A., FAYE, O., FAYE, O., MAGASSOUBA, N., WILLIAMS, C. V., AMBURGEY, V., WINONA, L., DAVIS, E., GERLACH, J., WASHINGTON, F., MONTEIL, V., JOURDAIN, M., BERERD, M., CAMARA, A., SOMLARE, H., CAMARA, A., GERARD, M., BADO, G., BAILLET, B.,

DELAUNE, D., NEBIE, K. Y., DIARRA, A., SAVANE, Y., PALLAWO, R. B., GUTIERREZ, G. J., MILHANO, N., ROGER, I., WILLIAMS, C. J., YATTARA, F., LEWANDOWSKI, K., TAYLOR, J., RACHWAL, P., TURNER, D., POLLAKIS, G., HISCOX, J. A., MATTHEWS, D. A., O'SHEA, M. K., JOHNSTON, A. M., WILSON, D., HUTLEY, E., SMIT, E., DI CARO, A., WOELFEL, R., STOECKER, K., FLEISCHMANN, E., GABRIEL, M., WELLER, S. A., KOIVOGUI, L., DIALLO, B., KEITA, S., RAMBAUT, A., FORMENTY, P., et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature,* 530**,** 228-232.

RAMSAY, J. P., KWONG, S. M., MURPHY, R. J., YUI ETO, K., PRICE, K. J., NGUYEN, Q. T., O'BRIEN, F. G., GRUBB, W. B., COOMBS, G. W. & FIRTH, N. 2016. An updated view of plasmid conjugation and mobilization in Staphylococcus. *Mob Genet Elements,* 6**,** e1208317.

READ, B. A., KEGEL, J., KLUTE, M. J., KUO, A., LEFEBVRE, S. C., MAUMUS, F., MAYER, C., MILLER, J., MONIER, A., SALAMOV, A., YOUNG, J., AGUILAR, M., CLAVERIE, J. M., FRICKENHAUS, S., GONZALEZ, K., HERMAN, E. K., LIN, Y. C., NAPIER, J., OGATA, H., SARNO, A. F., SHMUTZ, J., SCHROEDER, D., DE VARGAS, C., VERRET, F., VON DASSOW, P., VALENTIN, K., VAN DE PEER, Y., WHEELER, G., DACKS, J. B., DELWICHE, C. F., DYHRMAN, S. T., GLOCKNER, G., JOHN, U., RICHARDS, T., WORDEN, A. Z., ZHANG, X. & GRIGORIEV, I. V. 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature,* 499**,** 209-13.

REZNIKOFF, W. S. 2003. Tn5 as a model for understanding DNA transposition. *Mol Microbiol,* 47**,** 1199-206.

RICCI, G., BORGO, F. & FORTINA, M. G. 2006. Plasmids from Lactobacillus helveticus: distribution and diversity among natural isolates. *Lett Appl Microbiol,* 42**,** 254-8.

ROBERTS, R. J., CARNEIRO, M. O. & SCHATZ, M. C. 2013. The advantages of SMRT sequencing. *Genome Biol,* 14**,** 405.

RONAGHI, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res,* 11**,** 3-11.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-5.

SALVETTI, E., FONDI, M., FANI, R., TORRIANI, S. & FELIS, G. E. 2013. Evolution of lactic acid bacteria in the order Lactobacillales as depicted by analysis of glycolysis and pentose phosphate pathways. *Syst Appl Microbiol,* 36**,** 291-305.

SALVETTI, E., TORRIANI, S. & FELIS, G. E. 2012. The Genus Lactobacillus: A Taxonomic Update. *Probiotics Antimicrob Proteins,* 4**,** 217-26.

SALZBERG, S. L., DELCHER, A. L., KASIF, S. & WHITE, O. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res,* 26**,** 544-8.

SANCHEZ 2011. Introduction to Next Generation Sequencing.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1992. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology,* 24**,** 104-8.

SCHADT, E. E., TURNER, S. & KASARSKIS, A. 2010. A window into third-generation sequencing. *Hum Mol Genet,* 19**,** R227-40.

SEGERMAN, B. 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front Cell Infect Microbiol,* 2**,** 116.

SERENO, P. C. & LEE, M. 2005. The Logical Basis of Phylogenetic Taxonomy. *Systematic Biology,* 54**,** 595-619.

SHAO, Q., TRINH, J. T., MCINTOSH, C. S., CHRISTENSON, B., BALAZSI, G. & ZENG, L. 2017. Lysis-lysogeny coexistence: prophage integration during lytic development. *Microbiologyopen,* 6.

SIMS, D., SUDBERY, I., ILOTT, N. E., HEGER, A. & PONTING, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet,* 15**,** 121-32.

SLACK, J. 2014. *Genes: A Very Short Introduction*, Oxford University Press.

SMOKVINA, T., WELS, M., POLKA, J., CHERVAUX, C., BRISSE, S., BOEKHORST, J., VAN HYLCKAMA VLIEG, J. E. & SIEZEN, R. J. 2013. Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS One,* 8**,** e68731.

SNIEGOWSKI, P. D., GERRISH, P. J., JOHNSON, T. & SHAVER, A. 2000. The evolution of mutation rates: separating causes from consequences. *Bioessays,* 22**,** 1057-66.

STALDER, T. & TOP, E. 2016. Plasmid transfer in biofilms: a perspective on limitations and opportunities. *NPJ Biofilms Microbiomes,* 2.

STURTEVANT, A. H. 1913. A THIRD GROUP OF LINKED GENES IN DROSOPHILA AMPELOPHILA. *Science,* 37**,** 990-2.

TANIGUCHI, Y., CHOI, P. J., LI, G. W., CHEN, H., BABU, M., HEARN, J., EMILI, A. & XIE, X. S. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science,* 329**,** 533-8.

TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A. & KOONIN, E. V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res,* 28**,** 33-6.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A,* 102**,** 13950-5.

THOMAS, C. M. & SUMMERS, D. 2001. Bacterial Plasmids. *eLS.* John Wiley & Sons, Ltd.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-80.

TUOHIMAA, A., RIIPINEN, K. A., BRANDT, K. & ALATOSSAVA, T. 2006. The genome of the virulent phage Lc-Nu of probiotic Lactobacillus rhamnosus, and comparative genomics with Lactobacillus casei phages. *Arch Virol,* 151**,** 947-65.

TURING, A. 1936. On Computable Numbers. *Entscheidungsproblem*.

URSING, B. M. & ARNASON, U. 1998. Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. *Proc Biol Sci,* 265**,** 2251-5.

VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D., EISEN, J. A., WU, D., PAULSEN, I., NELSON, K. E., NELSON, W., FOUTS, D. E., LEVY, S., KNAP, A. H., LOMAS, M. W., NEALSON, K., WHITE, O., PETERSON, J., HOFFMAN, J., PARSONS, R., BADEN-TILLSON, H., PFANNKOCH, C., ROGERS, Y. H. & SMITH, H. O. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304**,** 66-74.

VENTURA, M., CANCHAYA, C., BERNINI, V., ALTERMANN, E., BARRANGOU, R., MCGRATH, S., CLAESSON, M. J., LI, Y., LEAHY, S., WALKER, C. D., ZINK, R., NEVIANI, E., STEELE, J., BROADBENT, J., KLAENHAMMER, T. R., FITZGERALD, G. F., O'TOOLE P, W. & VAN SINDEREN, D. 2006. Comparative genomics and transcriptional analysis of prophages identified in the genomes of Lactobacillus gasseri, Lactobacillus salivarius, and Lactobacillus casei. *Appl Environ Microbiol,* 72**,** 3130-46.

VENTURA, M., CANCHAYA, C., KLEEREBEZEM, M., DE VOS, W. M., SIEZEN, R. J. & BRUSSOW, H. 2003. The prophage sequences of Lactobacillus plantarum strain WCFS1. *Virology,* 316**,** 245-55.

VENTURA, M., CANCHAYA, C., PRIDMORE, R. D. & BRUSSOW, H. 2004. The prophages of Lactobacillus johnsonii NCC 533: comparative genomics and transcription analysis. *Virology,* 320**,** 229-42.

VOGEL, C., BASHTON, M., KERRISON, N. D., CHOTHIA, C. & TEICHMANN, S. A. 2004. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol,* 14**,** 208-16.

WANG, T. T. & LEE, B. H. 1997. Plasmids in Lactobacillus. *Crit Rev Biotechnol,* 17**,** 227-72.

WANG, Z., CHEN, Y. & LI, Y. 2004. A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics,* 2**,** 216-21.

WARD, N. & MORENO-HAGELSIEB, G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLoS One,* 9**,** e101850.

WATSON, J. D. 1970. *Molecular Biology of the Gene*, W. A. Benjamin.

WATSON, J. D. & CRICK, F. H. 1953. The structure of DNA. *Cold Spring Harb Symp Quant Biol,* 18**,** 123-31.

WEGMANN, U., MACKENZIE, D. A., ZHENG, J., GOESMANN, A., ROOS, S., SWARBRECK, D., WALTER, J., CROSSMAN, L. C. & JUGE, N. 2015. The pan-genome of Lactobacillus reuteri strains originating from the pig gastrointestinal tract. *BMC Genomics,* 16**,** 1023.

WETTERSTRAND 2012. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata.

WUYTS, S., WITTOUCK, S., DE BOECK, I., ALLONSIUS, C. N., PASOLLI, E., SEGATA, N. & LEBEER, S. 2017. Large-Scale Phylogenomics of the Lactobacillus casei Group Highlights Taxonomic Inconsistencies and Reveals Novel Clade-Associated Features. *mSystems,* 2.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol,* 11**,** 367-72.

YU, C., ZAVALJEVSKI, N., DESAI, V. & REIFMAN, J. 2011. QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Res,* 39**,** e88.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res,* 18**,** 821-9.

ZHANG, J. & YANG, J. R. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet,* 16**,** 409-20.

ZHENG, J., ZHAO, X., LIN, X. B. & GANZLE, M. 2015. Comparative genomics Lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. *Sci Rep,* 5**,** 18234

Chapter II

Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera

Contributions of thesis candidate:

1. Genome assembly and annotation that lead to all subsequent functional analyses.
2. Management of data storage and transfer among collaborators.
3. Analyses that contributed to all figures and tables except for Fig. 1, Supp. Fig. 4 and 5, Supp. Table 3 and 4.
4. Generation of Fig. 2 and 3, Supp. Fig. 1-3, 7, 9, 13-16, 19, 21-25, Supp. Table 1 and 2.
5. Writing of sections corresponding to the above figures and tables.
6. Review and submission of the final manuscript.
7. Responses and rebuttal of reviewers' comments for both re-submissions and relevant alteration of the manuscript.

Published as:

Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera.

Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Paul Ross R, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW.

Nat Commun. 2015 Sep 29; 6:8322.

doi: 10.1038/ncomms9322.

PMID: 26415554

TABLE OF CONTENTS

1 Introduction 77
2 Methods 79
2.1 Sequencing and assembly 79
2.2 CDS prediction and annotation 79
2.3 Construction of core- and pan-gene families 80
2.4 Assessing the robustness of core gene number and tree topology 80
2.5 Calculation of ANI and TNI 81
2.6 Phylogenetic analysis 81
2.7 Prediction of glycolysis-related genes 82
2.8 Bacteriocin prediction 82
2.9 Amino acid pathway identification 82
2.10 CRISPR identification 83
2.11 Investigation of niche association 83
2.12 Profiling of GHs and GTs 83
2.13 Identifying carbohydrate transporters 84
2.14 General metabolism 84
2.15 Identifying genes involved in stress response 84
2.16 Identification of Insertion Sequences 85
2.17 Phage identification 85
2.18 Plasmid identification 85
2.19 Analysis of LPXTG proteins, sortases and pilus gene clusters 85
2.20 Cell envelope protease (CEP) identification and analysis 86
3 Results 87
3.1 A genus more diverse than a family 87
3.2 A paraphyletic genus intermixed with five other genera 88
3.3 A broad repertoire of carbohydrate active enzymes 93
3.4 Sorting the interaction factors on the *Lactobacillus* cell surface 98
3.5 Differential evolution of Cell Envelope Protease genes 100
3.6 CRISPR-Cas systems and mobile genetic elements 101
4 Discussion 105
5 Bibliography 108

1 INTRODUCTION

The genus *Lactobacillus* comprises over 200 formally recognized species and subspecies that have been isolated from a wide range of sources (Salvetti et al., 2012). Their ability to ferment raw materials including milk, meat and plants has resulted in their industrial and artisanal use. Hence many *Lactobacillus* species have a long history of human usage (Bernardeau et al., 2006), including recognition as Generally Recognized as Safe (GRAS) or a Qualified Presumption of Safety (QPS) by FDA and EFSA, respectively (Bernardeau et al., 2008). Some strains are marketed as probiotics, meaning they may be beneficial to the consumer beyond basic nutritional value (Klaenhammer et al., 2012, Hill et al., 2014). Products containing lactobacilli dominate the global probiotics market, which is expected to reach a value of USD$24 billion by 2017. In addition to fermentative and preservative properties, some lactobacilli produce exopolysaccharides that contribute to the texture of foods (Badel et al., 2011), and to intestinal survival of probiotic species (Marco et al., 2010). Furthermore, lactobacilli are under development as delivery systems for vaccines (Mohamadzadeh et al., 2009) and therapeutics (Alvarez-Sieiro et al., 2014, Bermudez-Humaran et al., 2013). In recent years the relevance of lactobacilli to the chemical industry has considerably increased because of their capacity to produce enantiomers of lactic acid used for bioplastics as well as 1,3-propanediol (a starting ingredient used for biomedicines, cosmetics, adhesives, plastics and textiles) (Reddy et al., 2008). Thus, lactobacilli are among the microbes most commonly used for producing lactate from raw carbohydrates and synthetic media (Castillo Martinez et al., 2013).

The lactobacilli were originally grouped taxonomically according to their major carbohydrate metabolism, as homofermentative (metabolic group A), facultatively heterofermentative (group B) or obligately heterofermentative lactobacilli (group C) (Hammes and Vogel, 1995). The accumulation of 16S rRNA gene sequences (Collins et al., 1991) and a handful of genome sequences led to the realization that taxonomic and phylogenetic groupings of the lactobacilli were not concordant (Canchaya et al., 2006, Kant et al., 2011, Zhang et al., 2011, Makarova et al., 2006), that the genus is unusually diverse (as recently reviewed (Salvetti et al.,

2012)), and that a revised genome-based re-classification of the genus was warranted (Claesson et al., 2008).

To provide an extensive resource for comparing, grouping and functionally exploiting the lactobacilli, we sequenced 175 *Lactobacillus* genomes and 26 genomes from 8 other genera historically associated with or grouped within the lactobacilli. We complemented our analysis with the inclusion of 12 genome sequences from two genera that were already publicly available. In all but one case we sequenced genomes of Type Strains sourced from international culture collections (Supplementary Table 1), to provide taxonomic rigor and to avoid the problems associated with the genome sequence of a non-type strain unintentionally becoming the *de facto* genetic reference for that species, even when it contravened the published type-strain phenotype for that species (Felis et al., 2007). This phenomenon has added to confusion on strain identification. Three non-type strain *Leuconostoc* genomes were downloaded from NCBI (JB16, KM20 and 4882) and one *Pediococcus* non-type strain was sequenced (AS1.2696).

2 METHODS

2.1 SEQUENCING AND ASSEMBLY

Whole-genome sequencing was performed using Illumina HiSeq 2000 (Illumina Inc. U.S.A) by generating 100 bp paired-end read libraries following the manufacturer's instructions. An average of 190 Mb of high quality data were generated for each strain, corresponding to a sequencing depth of 16-fold to 185-fold (Supplementary Table 1).

The paired-end reads were first *de novo* assembled using SOAPdenovo v1.06, local inner gaps were then filled, and single base errors were corrected using the software GapCloser. The individual genome assemblies of 200 strains have been deposited in the National Center for Biotechnology Information under the project numbers PRJEB3060 and PRJNA222257 with individual accession numbers listed in Supplementary Table 1. Raw reads for 200 strains have been deposited in the sequence read archive (SRA) under the sample accession IDs listed in Supplementary Table 1.

2.2 CDS PREDICTION AND ANNOTATION

The coding sequences (CDS) of genes were predicted for each sequenced genome by using Glimmer v3.02 (Delcher et al., 2007). Partial genes were predicted by replacing gaps between contigs by a six-frame start/stop sequence (NNNNNCACACACTTAATTAATTAAGTGTGTGNNNNN). Glimmer3 normally predicts only complete genes, but a partial gene at a contig boundary with the above sequence at one or both ends will be predicted and given artificial end(s) (e.g. NNNNNCACACACTTAA at the 3' end). The number of partial genes along with their status (5' end missing, 3' end missing, both ends missing) were determined using these artificial ends. To obtain functional annotation, the amino acid sequences of predicted CDS were blasted (BLASTP) against the nr database with the criterion of e-value < 1e-5, identity > 40% and length coverage of gene > 50%. Additional annotation was obtained from the COG (Tatusov et al., 2003) and KEGG (Kanehisa et al., 2014) databases using BLASTP and the same BLAST thresholds.

2.3 CONSTRUCTION OF CORE- AND PAN-GENE FAMILIES

For identifying the pan-genome, a pair-wise comparison was performed using *L. gasseri* ATCC33323 as the first genome, followed by the random selection of each of the remaining genomes, without replacement, until all 213 genomes were included. Gene families were identified where homologous genes were found with BLASTP above the threshold of 25% identity over 40% of the gene length. Genes that fell below these thresholds formed new families, all of which were summed to give the pan-genome family set. A pan-genome family set was also derived after removing the genomes with greater than 200 contigs to assess the effect of higher contig number on pan-genome size.

To identify core genes for phylogenetic analysis, gene predictions for the 213 genomes were translated from nucleotide into amino acid sequences and used as the input for QuartetS (Zhang et al., 2011). QuartetS first predicted orthologs by reciprocal best BLAST between pairs of genomes using cut-offs of 25% identity and 40% length. The level of identity was kept above 25% given that below this level we cannot assume the shared common ancestry of genes based on sequence data alone (Chung and Subbiah, 1996). An equation that approximates the construction of a quartet gene tree assigned a confidence value to each reciprocal best blast pair of genes to determine if their relationship was orthologous or paralogous. Two-stage clustering (MCL and SLC) was used to cluster orthologs across all 213 genomes so that a presence and absence distribution could be determined for all gene families. Gene families with a representative sequence in all 213 genomes were selected as core genes for the construction of a phylogenetic tree. This method supported a core of 73 genes (Supplementary Table 2; Supplementary dataset S1 for sequences), which was used in all phylogenetic inferences.

2.4 ASSESSING THE ROBUSTNESS OF CORE GENE NUMBER AND TREE TOPOLOGY

We tested for the presence of 114 bacterial core marker genes (Wu et al., 2013) in the gene sequences of each of the 213 genomes and found that, while no genome had a low number of predicted marker genes (range 96 - 111), the 4

genomes with fewer than 105 genes all had contig numbers less than 200. Furthermore, when we correlated the number of predicted core genes (out of 114) with contig number, the Spearman correlation value was very low (rho value of 0.078; p-value = 0.26). This shows that draft genomes with larger contig numbers do not have artificially low core gene numbers.

To investigate the effect of core gene number on robustness of phylogeny, we omitted some of the more peripherally related LAB from the analysis, namely, we omitted the *Atopobium*, *Kandleria*, *Olsenella* and *Lactococcus* species, and this resulted in a core genome of 121 genes. The resulting phylogeny was highly congruent with the 73-core gene phylogeny, and was also supported by equally high bootstrap values. We put back in *Lactococcus* and removed *Carnobacterium*, resulting in a core gene set of 117 genes. Similarly, the resulting phylogeny was highly congruent with the 73-core gene phylogeny, and was also supported by equally high bootstrap values.

2.5 CALCULATION OF ANI AND TNI

The pair-wise ANI and TNI values across newly sequenced genomes were calculated according to methods proposed by Goris et al. (Goris et al., 2007) and Chen et al. (Chen et al., 2013), respectively. The frequency distributions of the ANI and TNI values of 3,730 published bacterial genomes were acquired from our previous report (Chen et al., 2013).

2.6 PHYLOGENETIC ANALYSIS

To determine the placement of the *Lactobacillus* Genus complex and associated genera within the Bacterial kingdom, we used AMPHORA2 (Wu and Scott, 2012), a marker gene database used in the phylogenetic inference of prokaryotes, to identity 16 marker genes (Supplementary Table 4; Dataset S2 for gene sequences), out of a total of 31 possible marker genes, that were shared across 452 representative bacterial species (Supplementary Table 3). We aligned the amino acid sequences for each gene separately using MUSCLE v3.8.31 (Edgar, 2004) and then constructed the maximum likelihood tree based on the concatenated alignment using the software PHYML with the WAG model (Guindon and Gascuel, 2003).

A Maximum Likelihood phylogeny concentrating on the *Lactobacillus* Genus complex and associated genera was inferred from 73 core genes present in all 213 strains. Amino acid sequences were aligned as above and the phylogeny was estimated using the PROTCATWAG model in RAxML v8.0.22 (Stamatakis, 2014) and rooted using *Atopobium minutum* DSM 20586, *Olsenella uli* DSM 7084 and *Atopobium rimae* DSM 7090. Bootstrapping was carried out using 100 replicates and values are indicated on the nodes of the phylogeny.

2.7 PREDICTION OF GLYCOLYSIS-RELATED GENES

A matrix with the presence/absence of the 10 core glycolytic genes across the 213 genomes was built using a combination of annotation querying and BLAST searching. When a gene was absent in one or more genomes, the result was confirmed with a tblastn (Altschul et al., 1990) search using *L. salivarius* query genes. In cases where a homolog was found using the blast approach the sequence was retrieved and aligned with mafft (Katoh and Toh, 2008). Alignments were inspected to confirm similarity of the sequences.

We mined the genomes for the presence of phosphoglycerate mutase using the approach published by Foster et al (Foster et al., 2010). The query phosphoglycerate mutases from *E. coli* GpmA (dPGM; NCBI GI number 50402115) and *E. coli* GpmM (iPGM,; 586733) were aligned against the six-frame translations of the 213 draft genomes with tblastn. Hits with a bit score larger than 100 were considered as a PGM match.

2.8 BACTERIOCIN PREDICTION

BAGEL (de Jong et al., 2010) was utilized to mine genomes for potential bacteriocin operons; results were manually verified within Artemis (Rutherford et al., 2000).

2.9 AMINO ACID PATHWAY IDENTIFICATION

Amino acid pathways were investigated through the KEGG suite of tools (Moriya et al., 2007).

2.10 CRISPR IDENTIFICATION

CRISPR-Cas systems were identified using CRISPRFinder (Grissa et al., 2007) and manual curation of the results.

2.11 INVESTIGATION OF NICHE ASSOCIATION

The 213 genomes were grouped into 6 niche categories in order to test for niche-specific associations in functional gene groups and genomic characteristics. The 6 niche categories are food (n=76), animal (n=56), plant (n=34), wine product (n=33), environment (n=7) and unknown (n=7). The niche category for each genome is shown in Table 1. We applied Kruskal-Wallis tests and generated boxplots for visualisation in order to determine trends among niches for 104 variables. These variables included all functional groups analysed in this study, MGEs (plasmids, phages and IS elements) and the following genomic parameters: genome size, gene number, contig number, GC content and sequencing depth. Statistics and visualisation were carried out in R v3.1.1.

2.12 PROFILING OF GHS AND GTS

The detection and assignment of sequences to families of carbohydrate-active enzymes (CAZyme) was carried out using a two-step approach. HMMSCAN (from the HMMER package v3.1b1) was used to query hidden Markov models representing the signature domains of each CAZyme family, to predict potential GTs and GHs across the 213 genomes below a threshold cut-off of 1e-05. In a separate approach, genes that have the GH and GT enzyme configuration (EC) designation EC 3.2.1.X and EC 2.4.X.X, respectively, were pooled into a GT and GH database. BLASTp searches were used to predict potential GTs and GHs from the 213 genomes using a cut off of 40% identity and 50% length with an e-value cut-off of 1e-05. Results from the HHM approach and the blast approach were compared to determine if both approaches supported the predicted gene results. Common genes

were retained and genes unique to one approach were screened against the Pfam 27.0 database to confirm the presence of GT/GH domains. Copy number of the verified GH/GT family were summarised in a heatmap.

2.13 IDENTIFYING CARBOHYDRATE TRANSPORTERS

To predict genes involved in carbohydrate transport we downloaded the protein database (go_20140614-seqdb.fasta.gz) from the Gene Ontology Consortium Database (http://archive.geneontology.org). A subset of this database was created by selecting all sequences that were annotated as carbohydrate transporters. Predicted genes from our study were blasted against this smaller database using BLASTP and genes involved in carbohydrate transport were selected using the thresholds, 40% identity, 50% coverage of query gene aligned and e-value <1e-05.

2.14 GENERAL METABOLISM

To generate an overview of metabolism we blasted all predicted genes against the STRING database v9 (Franceschini et al., 2013). The top hit for each gene (i.e. lowest e-value) was used to assign a COG category after applying thresholds of 40% identity, 50% of query gene length aligned and e-value <1e-05. R v3.1.1 was used for reformatting and for generating the COG heatmap.

2.15 IDENTIFYING GENES INVOLVED IN STRESS RESPONSE

The KEGG database was mined for gene products annotated as playing a part in stress responses. These were categorised into acid stress, oxidative stress, heat/DNA damage, cold stress, osmotic stress and bile tolerance. These genes were compiled into a database of 61,706 proteins. This database served to query (BLASTp) the predicted proteins encoded by the 213 genomes. Hits were considered stress response genes if their gene products displayed greater than 40% identity over 50% of the length of the KEGG stress response protein below an e-value of 1e-05. Copy number of the distribution of each of the stress-response proteins was summarised and visualised using a heat-map in the R statistical package v3.1.1.

2.16 IDENTIFICATION OF INSERTION SEQUENCES

To predict IS elements, Hidden Markov models representing 19 IS transposase families were downloaded from the TnpPred web service (http://www.mobilomics.cl). HMMSCAN (from the HMMER package v3.1b1) was used to query amino acid sequences of predicted genes against the HMMs.

2.17 PHAGE IDENTIFICATION

Bacteriophage genes were annotated by BLASTP search against the NCBI protein database using cut-offs of 40% identity over 50% of the length with an e-value of <1e-05. To predict phage-specific genes, a string search of predefined phage functions was carried out on gene annotations. Phage functions that overlap with non-phage functions such as those involved in transcription and DNA metabolism are usually annotated as belonging to prophages and these genes were also included in the phage results.

2.18 PLASMID IDENTIFICATION

For each genome, contigs were blasted against an NCBI reference database of complete plasmid sequences. A group of contigs was identified as belonging to a plasmid if at least 25% of their combined length aligned to at least 25% of the plasmid at ≥70% identity. These thresholds were determined empirically by adjusting alignment length and identity cut-offs until the strains in the dataset that are known to have plasmids and those that are known to have no plasmids both gave correct predictions. All predicted genes belonging to plasmid-associated contigs were then blasted against the STRING database v9.1 (Franceschini et al., 2013) in order to assign COG categories.

2.19 ANALYSIS OF LPXTG PROTEINS, SORTASES AND PILUS GENE CLUSTERS

Interproscan v. 5.44.0 with TIGRFAM 13.0 database with default parameters was used to search for conserved domains in the genomes (Haft et al., 2003,

Quevillon et al., 2005). Automatic pilus cluster search was performed using LOCP v. 1.0.0 with parameters "-P 1" and "-P_adj 0.05" (Plyusnin et al., 2009). The LOCP output results were then curated.. Both programs were run on the amino acid coding sequences data. R v. 3.0.1 was used for managing and parsing the output data (Team, 2013).

2.20 CELL ENVELOPE PROTEASE (CEP) IDENTIFICATION AND ANALYSIS

CEP sequences were identified in the genome sequences using two strategies. The first strategy involved a BLAST search using Subtilisin E as the search model. This returned 1,201 putative homologs. The second strategy used a HMM model for subtilisin as the search model and this returned 151 hits. Both panels of hits were further interrogated using the following strategy. Firstly, the presence of the key catalytic residues was confirmed (Asp, His and Ser, in this order of occurrence) and the proteins binned by number of residues in the sequence. The panels were further rationalized using a HMM search model for domains identified in the only solved structure of an active CEP, the ScpA from *Streptococcus pyogenes* (Kagawa et al., 2009). These searches included the DUF1034 which is equivalent to the Fn1 domain of ScpA, the CHU_C model corresponding to the Fn2 domain and the PA domain, SLAP which is an S layer anchoring domain and a manual inspection for LPXTG derivative sequence. This screening identified 60 CEPs across the genome database. Each of these hits was in turn used as a BLAST search model to confirm no additional CEPs could be identified. These searches proved to be internally consistent with no additional CEPs identified.

3 RESULTS

3.1 A GENUS MORE DIVERSE THAN A FAMILY

The genomes of the lactobacilli range in size from 1.23 Mb (*L. sanfranciscensis*) to four times larger (4.91 Mb; *L. parakefiri*) as shown in Supplementary Table 1 and Supplementary Fig. 1. The GC-content also varies considerably, from 31.93% to 57.02% (Supplementary Fig. 1). The core genome of the 213 strains comprises only 73 genes, the majority of which encode essential proteins for cell growth and replication (Supplementary Table 2). Owing to the draft nature of the genomes, this core gene number would increase were the genomes to be closed. The genus *Lactobacillus* and associated LAB genera have a large open pan-genome whose size increases continuously with the number of added genomes, and contains 44,668 gene families (Supplementary Fig. 2). Exclusion of draft genome assemblies at different fragmentation levels, namely greater than 20, 50, 100, 200, 300, 400 and 500 contigs does not lead to largely altered predictions for the pan-genome curves. Core genome curves were also generated using the same fragmentation levels and these curves are similar, especially for higher fragmentation levels. The core gene curves do show, however, that contig numbers have an effect on the core genome size (Supplementary Fig. 2). Although niche associations and described sources for *Lactobacillus* strains and species are not all equally robust, there was a clear trend for the genomes of species isolated from animals to be smaller, consistent with genome decay in a nutrient-rich environment (Makarova et al., 2006) (Supplementary. Fig 3).

ANI (average nucleotide identity) is the average identity value calculated from a pair-wise comparison of homologous sequences between two genomes and is frequently used in the definition of species (Chan et al., 2012, Goris et al., 2007). The frequency distribution of pair-wise ANI values for *Lactobacillus* species differs substantially from the distribution of values for Genus and Family, overlapping with values for Order and Class (Supplementary Fig. 4). TNI (total nucleotide identity) is an improved method that determines the proportion of matched nucleotide sequences between pairs of genomes, providing a higher discriminatory power for the high-level taxonomy units in this dataset (Chen et al., 2013). The TNI calculations

indicate that the genomic diversity of the genus *Lactobacillus* is intermediate between that of the majority of the currently approved taxonomic units for orders and families (http://www.bacterio.net/), and the mean value of total nucleotide identity between all species in this genus is 13.97% (Supplementary Fig. 4). Thus, although *Lactobacillus* has traditionally been defined as a Genus, its genetic diversity is larger than that of a typical Family.

3.2 A PARAPHYLETIC GENUS INTERMIXED WITH FIVE OTHER GENERA

In light of the extraordinary genomic diversity of the genus *Lactobacillus* and its polyphyletic nature, we set out to provide the most comprehensive phylogenetic study of the genus to date, thereby removing ambiguities in uncertain classifications and further validating existing taxonomic relationships. We constructed a phylogenetic tree with the lactobacilli and representative genomes of 452 selected genera from 26 phyla (Supplementary Table 3) using 16 proteins common to all taxa (see Supplementary Information for details and selection criteria; see Supplementary Table 4 for the protein list). The phylogeny revealed that *Lactobacillus* is paraphyletic and that all species of *Lactobacillus* descend from a common ancestor (Fig. 1; this tree with taxon names and branch lengths is presented in Supplementary Fig. 5). However, five other genera, *Pediococcus*, *Weissella*, *Leuconostoc*, *Oenococcus*, and *Fructobacillus*, are grouped within the lactobacilli as sub-clades. This phylogenomic arrangement was confirmed by a maximum likelihood tree constructed from the 73 core proteins shared by the 213 genomes of the lactobacilli and 10 associated genera (Fig. 2). This tree is supported by high boot-strap values, which supports the 73 core proteins as being truly reflective of the evolutionary history of the lactobacilli and associated genera, unbiased by HGT. The genera *Pediococcus*, *Leuconostoc* and *Oenococcus* have long been recognized as a phylogroup within the genus *Lactobacillus* based on both 16S rRNA gene sequence typing and extensive phylogenomic analysis (Makarova et al., 2006, Salvetti et al., 2012). Our results provide unequivocal evidence that the genera *Fructobacillus* and *Weissella* are members of the *Lactobacillus* clade, with *Fructobacillus* located between *Leuconostoc* and *Oenococcus* and the genus *Weissella* located as a sister branch (Fig. 2). As the *Lactobacillus* clade includes species from six different genera (*Lactobacillus*, *Pediococcus*, *Weissella*, *Leuconostoc*, *Oenococcus* and

Fructobacillus), we propose to name these six genera as constituting the *Lactobacillus* Genus Complex. Interestingly, the *Carnobacteria* are external to the *Streptococcus*/*Lactococcus* branch in the 16-core phylogeny of 26 phyla (Fig. 1), but they are internal to this branch in the 73-core tree of the *Lactobacillus* Genus Complex and associated genera (Fig. 2). The lower bootstrap values of 48% (*L. lactis*) and 64% (*Carnobacterium*) for the 16-core tree, which was built from an alignment of 3,863 bp, suggests that there was not enough phylogenetic signal to resolve these branches to a high degree of confidence. In contrast, the 73-core tree, which was built from an alignment of 30,780 bp, has bootstrap values of 100% for both these branches. This places greater confidence in the latter tree topology and hence it was used in all downstream analyses.

As a complement to the maximum likelihood tree of the *Lactobacillus* Genus Complex and associated genera based on 73 core proteins (Fig. 2), we built another tree (Supplementary Fig. 6) omitting *Atopobium*, *Olsenella*, *Kandleria* and *Carnobacterium* genomes and retaining the position of the most recent common ancestor (MRCA) according to the tree of bacteria (Fig. 2). In agreement with previous observations based on 28 LAB genomes (Zhang et al., 2011), this tree shows that the *Lactobacillus* Genus Complex splits into two main branches after diverging from the MRCA. Branch 1 contains the type species of the genus *Lactobacillus*, *L. delbrueckii*, and a large number of type strains that were isolated from dairy products. Branch 2 contains more species (n=127) than Branch 1 (n=77), and all five of the other genera in the *Lactobacillus* Genus Complex.

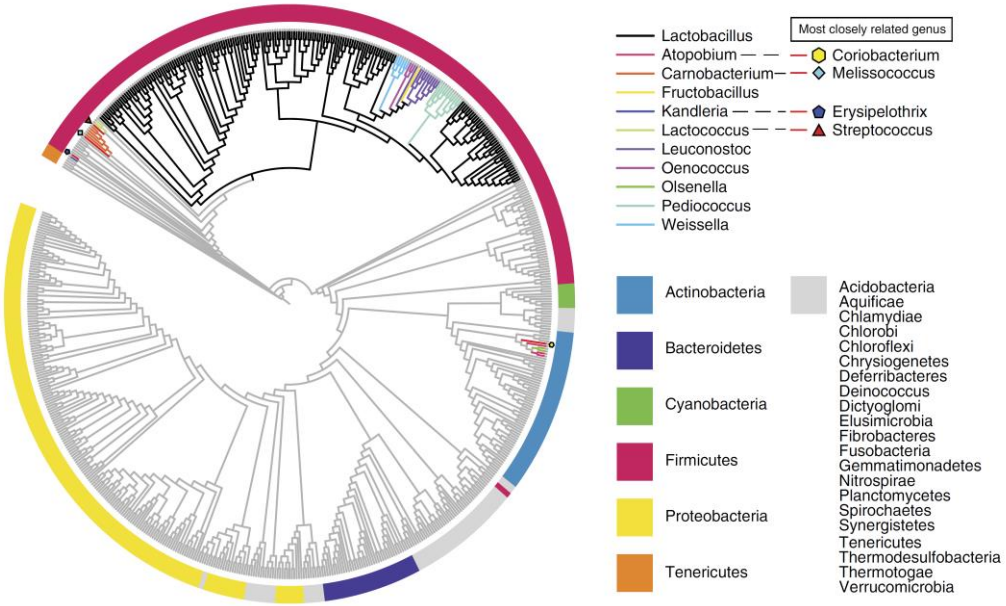


Figure 1: Cladogram of 452 genera from 26 phyla with the 213 genomes analysed in this study, based on the amino acid sequences of 16 marker genes. The tree was built by using the maximum likelihood method but visualized by removing the branch length information. The colored branches indicate different genera sequenced in this research; grey branches indicate members of genera whose genomes were previously sequenced. The outer circle color represent the phyla that are indicated in the legend, and the different shapes near tips indicate the position of genera that most closely related with *Atopobium*, *Carnobacterium*, *Kandleria*, and *Lactococcus*, separately.

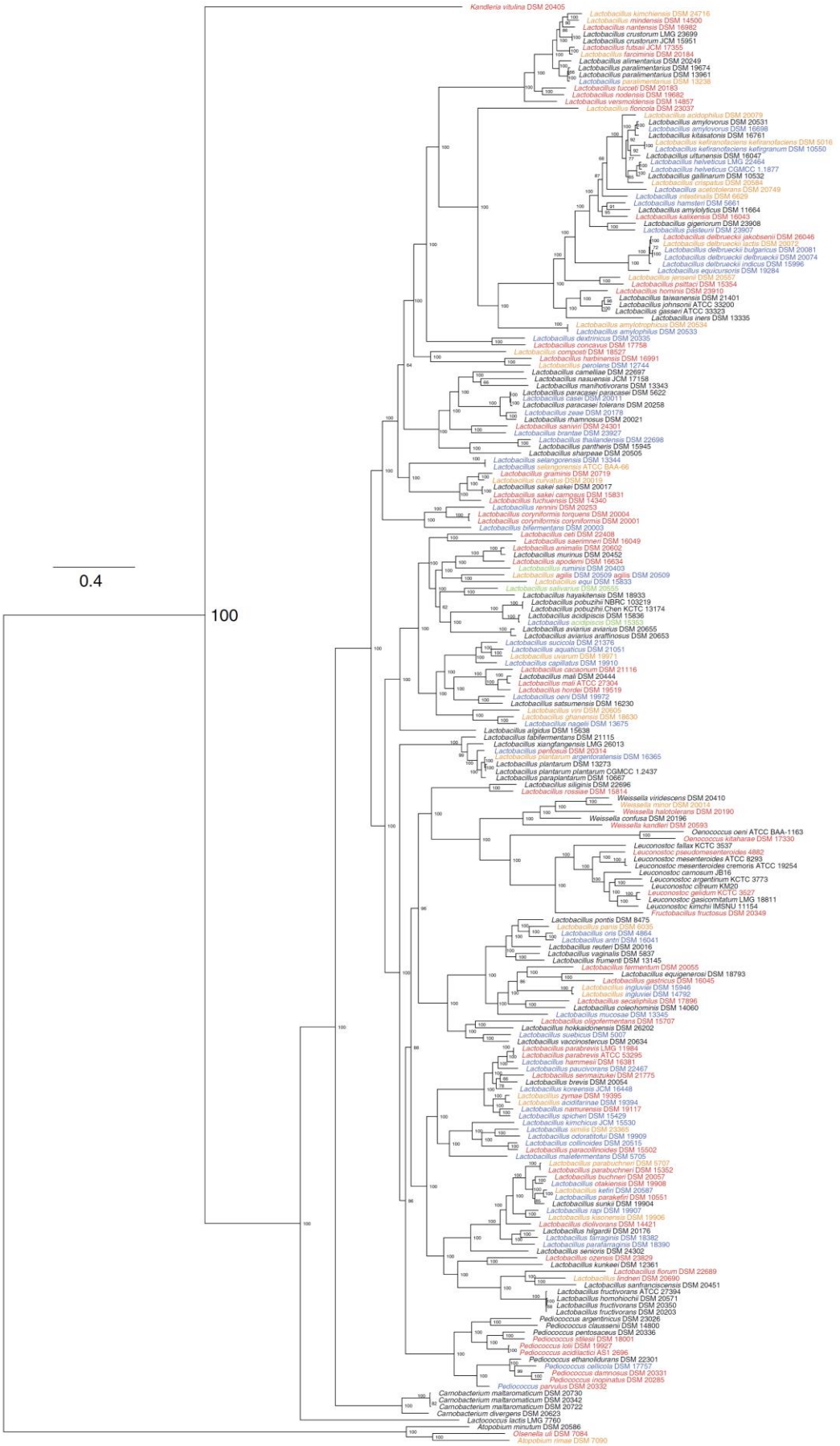

Kandleria vitulina DSM 20405
Lactobacillus kimchiensis DSM 24716
Lactobacillus mindensis DSM 14500
Lactobacillus nantensis DSM 16982
Lactobacillus crustorum LMG 23699
Lactobacillus crustorum JCM 15951
Lactobacillus futsaii JCM 17355
Lactobacillus farciminis DSM 20184
Lactobacillus alimentarius DSM 20249
Lactobacillus paralimentarius DSM 19674
Lactobacillus paralimentarius DSM 13961
Lactobacillus paralimentarius DSM 13238
Lactobacillus tucceti DSM 20183
Lactobacillus rodentium DSM 19682
Lactobacillus versmoldensis DSM 14857
Lactobacillus floricola DSM 23037
Lactobacillus acidophilus DSM 20079
Lactobacillus amylovorus DSM 20531
Lactobacillus amylovorus DSM 16698
Lactobacillus kitasatonis DSM 16761
Lactobacillus kefiranofaciens kefiranofaciens DSM 5016
Lactobacillus kefiranofaciens kefirgranum DSM 10550
Lactobacillus ultunensis DSM 16047
Lactobacillus helveticus LMG 22464
Lactobacillus helveticus CGMCC 1.1877
Lactobacillus gallinarum DSM 10532
Lactobacillus crispatus DSM 20584
Lactobacillus acetotolerans DSM 20749
Lactobacillus intestinalis DSM 6629
Lactobacillus hamsteri DSM 5661
Lactobacillus amylolyticus DSM 11664
Lactobacillus kalixensis DSM 16043
Lactobacillus gigeriorum DSM 23908
Lactobacillus pasteurii DSM 23907
Lactobacillus delbrueckii jakobsenii DSM 26046
Lactobacillus delbrueckii lactis DSM 20072
Lactobacillus delbrueckii bulgaricus DSM 20081
Lactobacillus delbrueckii delbrueckii DSM 20074
Lactobacillus delbrueckii indicus DSM 15996
Lactobacillus equicursoris DSM 19284
Lactobacillus jensenii DSM 20557
Lactobacillus paviani DSM 21790
Lactobacillus helveticus DSM 20075
Lactobacillus pelitacti DSM 15354
Lactobacillus hominis DSM 23910
Lactobacillus taiwanensis DSM 21401
Lactobacillus johnsonii ATCC 33200
Lactobacillus gasseri ATCC 33323
Lactobacillus iners DSM 13335
Lactobacillus amylotrophicus DSM 20534
Lactobacillus amylophilus DSM 20533
Lactobacillus dextrinicus DSM 20335
Lactobacillus concavus DSM 17758
Lactobacillus composti DSM 18527
Lactobacillus harbinensis DSM 16991
Lactobacillus perolens DSM 12744
Lactobacillus camelliae DSM 22697
Lactobacillus nasuensis JCM 17158
Lactobacillus manihotivorans DSM 13343
Lactobacillus paracasei paracasei DSM 5622
Lactobacillus casei DSM 20011
Lactobacillus paracasei tolerans DSM 20258
Lactobacillus zeae DSM 20178
Lactobacillus rhamnosus DSM 20021
Lactobacillus sanniviri DSM 24301
Lactobacillus brantae DSM 23927
Lactobacillus thailandensis DSM 22698
Lactobacillus pantheris DSM 15945
Lactobacillus sharpeae DSM 20505
Lactobacillus selangorensis DSM 13344
Lactobacillus selangorensis ATCC BAA-66
Lactobacillus graminis DSM 20719
Lactobacillus curvatus DSM 20019
Lactobacillus sakei sakei DSM 20017
Lactobacillus sakei carnosus DSM 15831
Lactobacillus fuchuensis DSM 14340
Lactobacillus rennini DSM 20253
Lactobacillus coryniformis torquens DSM 20004
Lactobacillus coryniformis coryniformis DSM 20001
Lactobacillus bifermentans DSM 20003
Lactobacillus ceti DSM 22408
Lactobacillus saerimneri DSM 16049
Lactobacillus animalis DSM 20602
Lactobacillus murinus DSM 20452
Lactobacillus apodemi DSM 16634
Lactobacillus ruminis DSM 20403
Lactobacillus agilis DSM 20509 agilis DSM 20509
Lactobacillus equi DSM 15833
Lactobacillus salivarius DSM 20555
Lactobacillus hayakitensis DSM 18933
Lactobacillus pobuzihi NBRC 103219
Lactobacillus pobuzihi Chen KCTC 13174
Lactobacillus acidipiscis DSM 15836
Lactobacillus acidipiscis DSM 15555
Lactobacillus avarius avarius DSM 20655
Lactobacillus avarius araffinosus DSM 20653
Lactobacillus aucocola DSM 21376
Lactobacillus aquaticus DSM 21051
Lactobacillus uvarum DSM 19971
Lactobacillus capillatus DSM 19910
Lactobacillus cacaonum DSM 21116
Lactobacillus mali DSM 20444
Lactobacillus mali ATCC 27304
Lactobacillus hordei DSM 19519
Lactobacillus oeni DSM 19972
Lactobacillus satsumensis DSM 16230
Lactobacillus viri DSM 20605
Lactobacillus ghanensis DSM 18630
Lactobacillus nagelii DSM 13675
Lactobacillus algidus DSM 15838
Lactobacillus fabifermentans DSM 21115
Lactobacillus xiangfangensis LMG 26013
Lactobacillus perfosus DSM 20314
Lactobacillus plantarum argentoratensis DSM 16365
Lactobacillus plantarum DSM 13273
Lactobacillus plantarum plantarum CGMCC 1.2437
Lactobacillus paraplantarum DSM 10667
Lactobacillus siliginis DSM 22696
Lactobacillus rossiae DSM 15814
Weissella viridescens DSM 20410
Weissella minor DSM 20014
Weissella halotolerans DSM 20190
Weissella confusa DSM 20196
Weissella kandleri DSM 20593
Oenococcus oeni ATCC BAA-1163
Oenococcus kitaharae DSM 17330
Leuconostoc fallax KCTC 3537
Leuconostoc pseudomesenteroides 4882
Leuconostoc mesenteroides ATCC 8293
Leuconostoc mesenteroides cremoris ATCC 19254
Leuconostoc carnosum JB16
Leuconostoc argentinum KCTC 3773
Leuconostoc citreum KM20
Leuconostoc gelidum KCTC 3527
Leuconostoc gasicomitatum LMG 18811
Leuconostoc kimchii IMSNU 11154
Fructobacillus fructosus DSM 20349
Lactobacillus pontis DSM 8475
Lactobacillus panis DSM 6035
Lactobacillus oris DSM 4864
Lactobacillus antri DSM 16041
Lactobacillus reuteri DSM 20016
Lactobacillus vaginalis DSM 5837
Lactobacillus frumenti DSM 13145
Lactobacillus fermentum DSM 20055
Lactobacillus equigenerosi DSM 18793
Lactobacillus gastricus DSM 16045
Lactobacillus ingluviei DSM 15946
Lactobacillus ingluviei DSM 14792
Lactobacillus secaliphilus DSM 17896
Lactobacillus coleohominis DSM 14060
Lactobacillus mucosae DSM 13345
Lactobacillus oligofermentans DSM 15707
Lactobacillus hokkaidonensis DSM 26202
Lactobacillus suebicus DSM 5007
Lactobacillus vaccinostercus DSM 20634
Lactobacillus parabrevis LMG 11984
Lactobacillus parabrevis ATCC 53295
Lactobacillus hammesii DSM 16381
Lactobacillus paucivorans DSM 22467
Lactobacillus senmaizukei DSM 21775
Lactobacillus brevis DSM 20054
Lactobacillus koreensis JCM 16448
Lactobacillus zymae DSM 19395
Lactobacillus acidifarinae DSM 19394
Lactobacillus namurensis DSM 19117
Lactobacillus spicheri DSM 15429
Lactobacillus kimchicus JCM 15530
Lactobacillus similis DSM 20482
Lactobacillus odoratitofui DSM 19909
Lactobacillus collinoides DSM 20515
Lactobacillus paracollinoides DSM 15502
Lactobacillus malefermentans DSM 5705
Lactobacillus parabuchneri DSM 5707
Lactobacillus parabuchneri DSM 15352
Lactobacillus buchneri DSM 20057
Lactobacillus otakiensis DSM 19908
Lactobacillus kefiri DSM 20587
Lactobacillus parakefiri DSM 10551
Lactobacillus sunkii DSM 19904
Lactobacillus rapi DSM 19907
Lactobacillus kisonensis DSM 19906
Lactobacillus diolivorans DSM 14421
Lactobacillus hilgardii DSM 20176
Lactobacillus farragine DSM 18382
Lactobacillus parafarraginis DSM 18390
Lactobacillus senioris DSM 24302
Lactobacillus ozensis DSM 23829
Lactobacillus kunkeei DSM 12361
Lactobacillus florum DSM 22689
Lactobacillus lindneri DSM 20690
Lactobacillus sanfranciscensis DSM 20451
Lactobacillus fructivorans ATCC 27394
Lactobacillus homohiochii DSM 20571
Lactobacillus fructivorans DSM 20350
Lactobacillus fructivorans DSM 20203
Pediococcus argentinicus DSM 23026
Pediococcus claussenii DSM 14800
Pediococcus pentosaceus DSM 20336
Pediococcus stilesii DSM 18001
Pediococcus lolii DSM 19927
Pediococcus acidilactici AS1 2696
Pediococcus ethanoliduran DSM 22301
Pediococcus cellicola DSM 17757
Pediococcus damnosus DSM 20331
Pediococcus inopinatus DSM 20285
Pediococcus parvulus DSM 20332
Carnobacterium maltaromaticum DSM 20730
Carnobacterium maltaromaticum DSM 20342
Carnobacterium maltaromaticum DSM 20722
Carnobacterium divergens DSM 20623
Lactococcus lactis LMG 7760
Atopobium minutum DSM 20586
Olsenella uli DSM 7084
Atopobium rimae DSM 7090
0.4
100

Figure 2: Maximum likelihood phylogeny derived from 73 core genes across 213 strains. The phylogeny was estimated using the PROTCATWAG model in RAxML and rooted using the branch leading to *Atopobium minutum* DSM 20586, *Olsenella uli* DSM 7084 and *Atopobium rimae* DSM 7090 as the outgroup. Bootstrapping was carried out using 100 replicates and values are indicated on the nodes. Colours on taxon labels indicated presence of CRISPR-Cas systems using pink, blue and green for Type I, II and III systems, respectively. Undefined systems are represented in yellow. Color combinations were used when multiple systems from different families were concurrently detected in bacterial genomes.

3.3 A BROAD REPERTOIRE OF CARBOHYDRATE ACTIVE ENZYMES

With interest in their applications in fermentations, some of the earliest classifications of lactobacilli were based on their carbohydrate utilization patterns (Hammes and Vogel, 1995). Glycolysis occurs in obligately homofermentative (group A) and facultatively heterofermentative (group B) lactobacilli, and has been traditionally linked to the presence of 1,6-biphosphate aldolase (Kandler, 1983). A full set of glycolysis genes were predicted in 49% of the species analysed (Supplementary Fig. 7), and gene duplication is common, though not particularly associated with a group or niche. All *Lactobacillus*, *Leuconostoc*, *Weissella*, *Fructobacillus* and *Oenococcus* species lacking phosphofructokinase (Pfk) formed a distinct monophyletic group. This group included the historically-defined *L. reuteri*, *L. brevis*, *L. buchneri*, *L. collinoides*, *L. vaccinostercus* and *L. fructivorans* groups. Most species (75%) within this Pfk-negative clade also lacked 1,6-biphosphate aldolase, though this gene was consistently present in the *Weissella* clade as well as in some leuconostocs and species from the *L. reuteri* and *L. fructivorans* groups. Importantly, most species (87%) within the Pfk-lacking group were classified as obligatively heterofermentative (Salvetti et al., 2012), with the rest being facultatively heterofermentative. The reason for the link between *pfk* gene loss and heterofermentative metabolism needs functional genomic investigation. The average phylogenetic distance (number of nodes to root) of facultatively heterofermentative lactobacilli (as defined in Supplementary Fig. 7) to the MRCA (Supplementary Fig. 6) is considerably lower than that of obligately heterofermentative or obligately homofermentative species (Supplementary Fig. 8) suggesting that the *Lactobacillus* MRCA was facultatively heterofermentative. The obligatively heterofermentative species also form a distinct cluster that may be explained by several evolutionary scenarios that require further investigation.

Biotransformation of carbohydrates by bacteria can be exploited for transforming raw materials, for optimizing growth and for producing valuable metabolites. The 213 genomes collectively encode 48 of the 133 families of glycoside hydrolases (GH) in the CAZy database (http://www.cazy.org), many of which represent unrecognized and unexploited enzymes for biotechnology (Fig. 3). Chitin is the second most abundant natural polysaccharide after cellulose. Among 115 LAB species previously tested, only *Carnobacterium* spp. were able to

hydrolyse alpha chitin (Leisner et al., 2008). In this study, three new *Carnobacterium* genomes, along with strains of *L. delbrueckii*, *L. nasuensis*, *L. agilis*, *L. fabifermentans* and *Pediococcus*, provide the genetic information to exploit that activity. The GH39 genes are beta-xylosidases that are present in the *L. rapi*/*L. kisonensis* branch as well as two singleton species, *L. concavus* and *L. secaliphilus*. GH49 (dextranase) and GH95 (alpha-fucosidase) are harboured only in the *L. harbinensis*/*L. perolens* branch with GH49 being absent from the latter species. Dextranases are considered to be the most efficient means for hydrolysing undesirable dextrans at sugar mills (Rodríguez Jiménez, 2009). Microbial mannanases hydrolyse complex plant polysaccharides and they have applications in the paper and pulp industry, for food and feed technology, coffee extraction, oil drilling and detergent production; the corresponding GH76 is found only in the two *L. acidipiscis* strains. GH101 is found only in *L. brantae* isolated from goose feces and *L. perolens* which is from a beverage production environment. This GH is an endo-alpha-N-acetylgalactosaminidase, which is thought to play a role in the degradation and utilization of mucins by probiotic bifidobacteria (Fujita et al., 2005). While this explains its presence in the goose intestine, its association with beverage production may be due to limited hygiene.

We identified two GH families not previously associated with the *Lactobacillus* genus complex. GH67 displays alpha-glucuronidase activity (Shallom et al., 2004) and is involved in the breakdown of xylan; such enzymes have an application in the pulp industry for bio-bleaching, in the paper industry, as food additives in poultry and in wheat flour for improving dough handling (Beg et al., 2001). GH95 fucosidases can cleave and remove specific fucosyl residues (Katayama et al., 2004). Fucose residues are present in oligosaccharides in milk and on erythrocyte surface antigens. Some GH types appeared to be common across the genome dataset, if not universal, and these are described in Supplementary Information.

Analysis of the 213 genomes reveals they encode representatives of 22 of the 95 families of glycosyltransferases (GT) in the CAZy database with a high level of GT-encoding diversity and a number of surprising findings (Supplementary Fig. 9). Glycogen is one of five main carbohydrate storage forms used by bacteria, and a previous analysis of 1,202 diverse bacteria concluded that bacteria that can synthesize glycogen occupy more diverse niches (Wang and Wise, 2011). GT5 and

GT35 are glycogen synthase and glycogen phosphorlyase respectively. These GTs are encoded by the *L. casei* clade, which includes two species that are currently exploited heavily as probiotics, *L. casei* and *L. rhamnosus*, as well as the *L. plantarum* group, some members of the *L. salivarius* group (such as *L. salivarius* itself) and a number of singletons. It is not clear if the ability to synthesize glycogen contributes to the biological fitness of these species. Strikingly, among the sequenced genomes only *L. gasseri* encodes GT11 (galactoside α-1,2-L-fucosyltransferase) while only *L. delbrueckii* DSM15996 encodes GT92 (N-glycan core α-1,6-fucoside β-1,4-galactosyltransferase). Surface fucose is common in pathogens, including *Helicobacter pylori*, where it is linked to antigenic mimicry (with Lewis blood group antigens), immune avoidance and adhesion (Bergman et al., 2006). According to the CAZy database, the GT11 fucosyltransferase is uncommon in LAB; it is present in *Akkermansia muciniphila*, in a minority of commensal *Bacteroides*, in three *Roseburia* species and in several Proteobacteria. Interestingly, GT92 is not described in any prokaryotic organisms in CAZy, but the current study identified the characteristic GT92 domain in *L. delbreuckii*. The production of surface fucose-containing moieties by certain *L. gasseri* and *L. delbreuckii* strains merits biological evaluation.

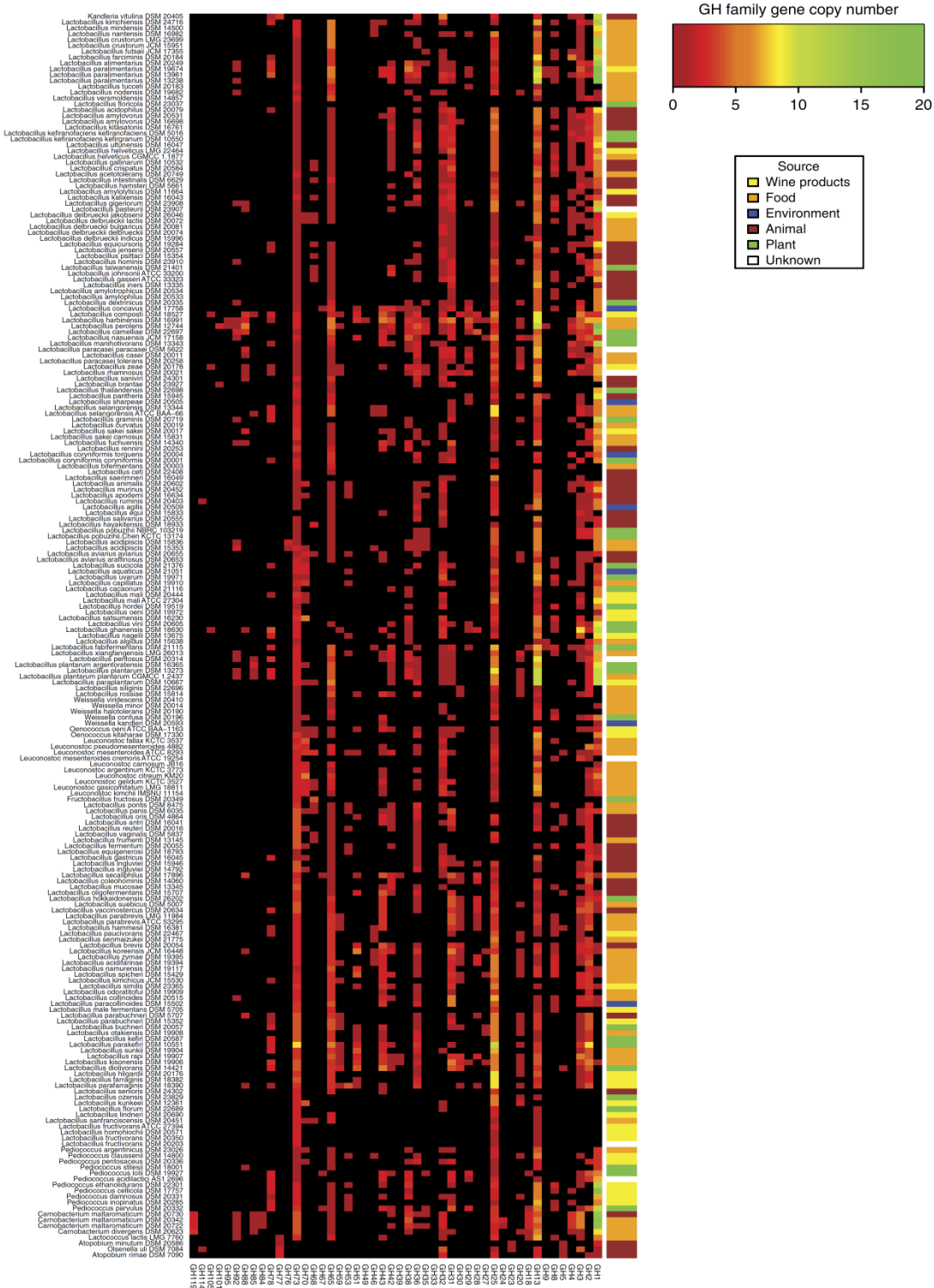

GH family gene copy number
0
5
10
15
20
Source
Wine products
Food
Environment
Animal
Plant
Unknown

Figure 3: Heatmap illustrating the distribution and abundance of glycoside hydrolase (GH) family genes across the *Lactobacillus* Genus Complex and associated genera. Gene copy number of each of the 48 represented GH families is indicated by the colour key ranging from black (absent) to green. Strains are graphed in the same order left to right as they appear top to bottom in the phylogeny (Fig. 2) with the isolation source of each strain indicated by the colour bar at the top of the heatmap.

3.4 SORTING THE INTERACTION FACTORS ON THE *LACTOBACILLUS* CELL SURFACE

Surface proteins of lactobacilli include key interaction receptors for probiotics and enzymes for growth in milk. A major class of surface proteins in Gram-positive bacteria are those anchored by sortase enzymes that recognize a highly conserved LPXTG sequence motif (Navarre and Schneewind, 1999). We identified 1,628 predicted LPXTG-containing proteins and 357 sortase enzymes in the 213 genomes (Supplementary Table 5). The number of sortases and LPXTG proteins greatly varies between species (Fig. 4), with 0 to 27 LPXTG proteins found. The highest number of LPXTG proteins (27) occurred in the milk isolate *Carnobacterium maltaromaticum* DSM 20342. Other species of the genus *Carnobacterium* also showed a large LPXTG protein repertoire, suggesting extensive interactions within their respective habitats and associated microbial communities. Among the variety of LPXTG proteins, we particularly focused on sortase-dependent pilus gene clusters. Common in Gram-positive pathogens, these proteinaceous fibers are also produced by commensal bacterial species such as *Lactobacillus rhamnosus* (Kankainen et al., 2009) and the SpaCBA pili have been shown to contribute to probiotic properties by mucin binding (von Ossowski et al., 2010) and cellular signalling (Ardita et al., 2014). A total of 67 pilus gene clusters were predicted in 51 bacterial strains (Fig. 4), most strains harboring a single pilus gene cluster (PGC) (Supplementary Fig. 10). Only about one third of the piliated strains possessed pilus gene clusters similar to *L. rhamnosus* strain GG pilus clusters in terms of gene order, i.e. a cluster of three pilin genes and one pilin-specific sortase gene. The remaining pilus clusters showed the presence of two other major types and numerous other types that are different in organization and sequence from that of *L. rhamnosus* GG (Fig. 4). Five particular clades were associated with the presence of PGCs. The ecologically diverse *L. casei*/*L. rhamnosus* clade (Figure 4, Panel C, Clade ii) harbored the greatest number of piliated species. Some strains e.g. *L. equicursoris*, *W. confusa* and *L. parabuchneri* (DSM 15352) are distinguished by being the only piliated species within their respective clades (Fig. 4, Panel C), which we cannot currently explain. The availability from this study of over 50 new pilus gene clusters is expected to provide new avenues for addressing their role in probiotic and other functions.

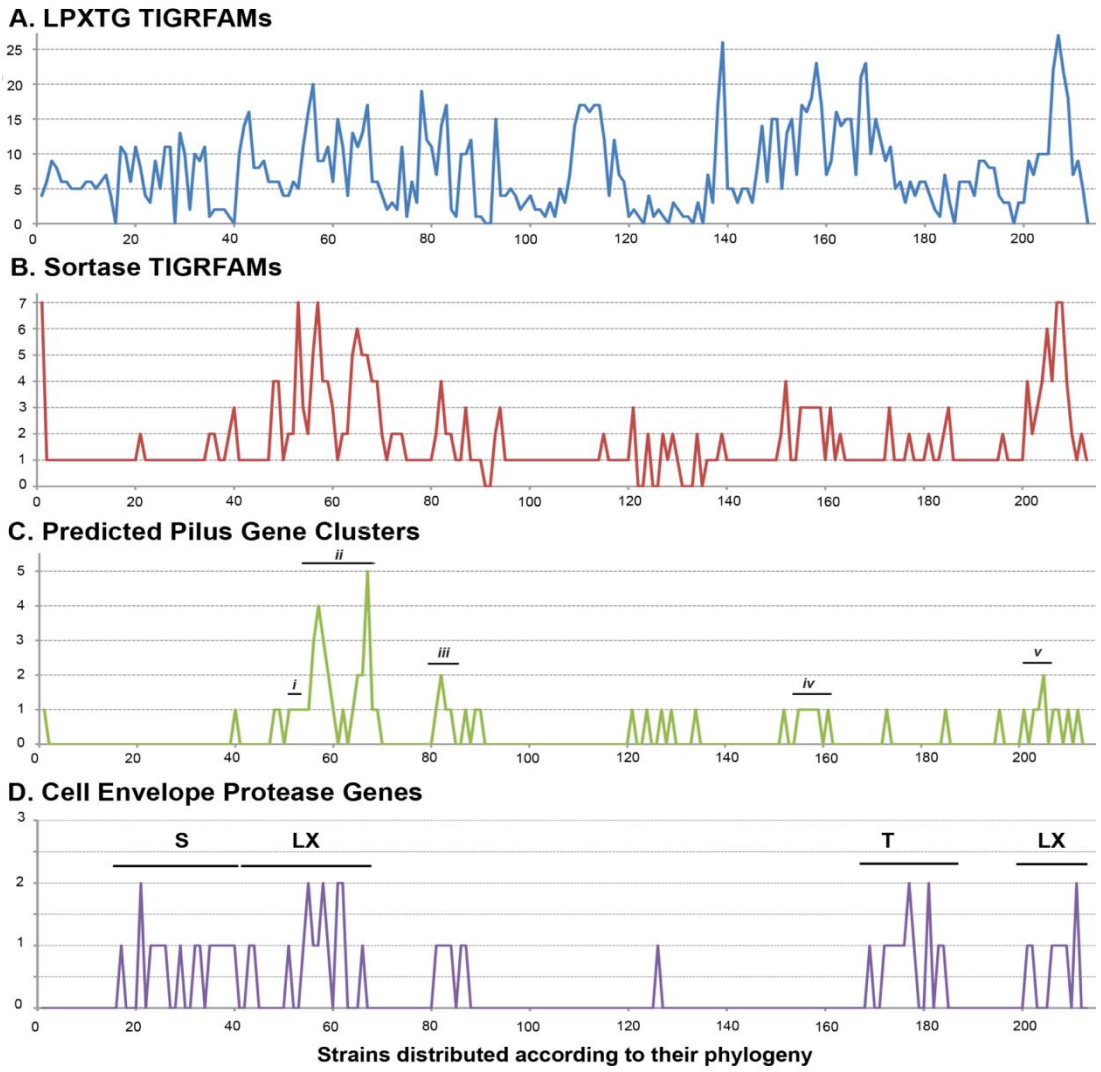


Figure 4: Differential abundance of genes encoding LPXTG proteins, sortases, pili and cell envelope proteases (Panels A, B, C and D, respectively). The y-axis indicates the number of genes/clusters detected. Strains are graphed in the same order left to right as they appear top to bottom in the phylogeny (Fig. 2). In panel C, each black bar indicates strains belonging to the same lineages. Panel C legend: *i*. the *L. composti* clade; *ii*. the *L. casei/rhamnosus* clade; *iii*. the *L. ruminis* clade; *iv*. the *L. brevis/parabrevis* clade; *v*. the *Pediococcus ethanolidurans* clade. Panel D legend: S, S-layer type anchor; LX, LPXTG-sortase dependent anchor (including derivatives thereof); T, truncated protein.

3.5 DIFFERENTIAL EVOLUTION OF CELL ENVELOPE PROTEASE GENES

Cell Envelope Proteases (CEP) are multi-subunit, cell-wall-anchored, subtilase-type proteinases produced by many LAB. They are primarily associated with cleaving casein as the first stage in releasing peptides and amino acids during growth in milk, and variations in their sequence and domain structure contribute to determining the flavour of cheese (Siezen, 1999). In particular, the Protease Associated (PA) domain and the A domain have been shown to impact on the specificity of the enzyme. The A domain has been subdivided into 3 fibronectin domains (Fn1, Fn2 and Fn3) and these are implicated in substrate binding (Kagawa et al., 2009). Furthermore some CEPs of commensal lactobacilli may act upon inflammatory mediators to ameliorate Inflammatory Bowel Disease (von Schillde et al., 2012), so mining the novel *Lactobacillus* genomes for these proteases could identify novel therapeutics for chemokine-mediated inflammatory diseases. We identified genes for 60 CEPs in the 213 genomes, ranging from 1,097 to 2,270 amino acids in length (Supplementary Table 6). Forty four strains had a single CEP, while 8 strains encoded 2 distinct CEPs (Fig. 4). Four disrupted CEP genes were detected, two occurring at contig boundaries. Presence of genes for CEPs exhibited clear clade association, notably with the *L. delbrueckii*, *L. casei* and *L. buchneri* clades, part of the *L. salivarius* clade, and the *Carnobacterium* clade.

The CEPs are defined as cell associated, and different anchoring mechanisms have been identified. Seventeen of the 60 CEPs incorporated a SLAP domain, putatively responsible for non-covalent interactions with the cell wall, 12 had a canonical LPXTG motif for covalent linkage to peptidoglycan, and a further 18 had a derivative of the LPXTG motif (Fig. 4). Interestingly, 13 of the CEPs had neither an S-layer type domain nor an LPXTG type motif. These proteins all terminated precisely before standard anchoring motifs at a sequence conserved across all of the 60 identified CEPs, suggesting that this was non-random. Of these 13 CEPs, 11 are in the *L. buchneri* clade, suggesting positive selection for release of protease activity into the growth medium in this clade. There may be an advantage to the cell by releasing enzyme away from the cell surface and not saturating or competing for cell wall anchoring. Twelve of these 13 CEPs cluster in a distinct group in a phylogenetic tree and the multiple alignment indicates the sequences differ from other CEPs along the entire length of the protein (data not shown). Putative anchoring by the SLAP

domain is notably associated with the *L. delbrueckii* sub-clade, while CEPs containing LPXTG motifs occur in the *L. casei*, *L. salivarius*, *Pediococcus* and *Carnobacterium* groups.

The pair-wise amino acid identity values between the 60 CEPS ranged from 100% down to just 20%, a level of divergence indicating the likelihood that some of these proteases have novel specificity. Of the 60 CEPs identified, 23 had the PA domain, 57 the Fn1 domain (DUF_1034) and 25 the Fn2 domain (CHU_C). Interestingly, there is some association between anchoring mechanism and domain composition. For the SLAP domain-containing CEPs, 12/17 do not contain the Fn2 domain, and for the CEPs devoid of SLAP or LPXTG sequences, 11/13 do not contain a PA domain. The differential domain composition in the CEPs indicates that a diverse range of substrates and products are likely. These properties may be exploitable for improvement of food flavour or for enhanced probiotic capabilities.

3.6 CRISPR-CAS SYSTEMS AND MOBILE GENETIC ELEMENTS

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) in combination with CRISPR-associated proteins (Cas) constitute CRISPR-Cas systems, which provide adaptive immunity against invasive elements in bacteria (Barrangou et al., 2007). Sequences derived from exogenic elements are integrated into CRISPR loci, transcribed and processed into mature small interfering RNAs, and the small CRISPR RNAs (crRNAs) specifically guide Cas effector proteins for sequence-dependent targeting and endonucleolytic cleavage of DNA sequences complementary to the spacer sequence (Barrangou and Marraffini, 2014). CRISPR-Cas systems have revolutionized genetic engineering and gene therapy by enabling precise targeted manipulations in prokaryotic (Jiang et al., 2013) and eukaryotic genomes (Hill et al., 2014), and recently in lactobacilli (Oh and van Pijkeren, 2014).

A total of 137 CRISPR loci were identified in 62.9% of the genomes analysed, representing all the major phylogenetic groups of lactobacilli evaluated (Fig. 2). This indicates that these systems are evolutionarily widespread throughout this genus, and likely functionally important. This is considerably higher than the ~46% general occurrence rate in bacterial genomes in CRISPRdb (Grissa et al., 2007). There was overall congruence between the phylogenomic structure of the lactobacilli (Fig. 2) and CRISPR-Cas system phylogeny (Supplementary Fig. 11)

reflecting co-evolutionary patterns. For Type allocation, the signature genes *cas3*, *cas9* and *cas10* for Types I, II and III, respectively, were used, complemented by comparison of CRISPR repeat sequences and the universal Cas1 protein (Grissa et al., 2007). Types I, II and III CRISPR-Cas systems were all detected (66, 68, and 3 systems, respectively; Supplementary Table 7). Comparative analyses of defining CRISPR features revealed a diversity of the universal Cas1 protein and corresponding CRISPR repeat sequences, with consistent clustering in two main families representing Type I and Type II systems (Supplementary Fig. 11). Strikingly, Type II systems were detected in 36% of the *Lactobacillus* Genus Complex and associated genera, though they occur in only 5% of all bacterial genomes analyzed to date (Chylinski et al., 2014), suggesting these LAB are a rich resource for Type II CRISPR systems. Beyond the diversity of CRISPR-Cas systems, we further uncovered dramatic variability in locus size and spacer content, ranging from 2 to 135 CRISPR spacers (Supplementary Table 7).

Type II CRISPR-Cas systems, which comprise the signature Cas9 endonuclease have received tremendous interest given their ability to re-program Cas9 using customized guide RNAs for sequence-specific genesis of double stranded breaks and the corresponding ability to edit genomes using DNA repair machinery. Here, we observed a diversity of novel Type II systems with heterogeneous Cas9 sequences (Supplementary Fig. 12, panel A) that expands the Cas9 space considerably, and the corresponding DNA targeting and cleavage features including the proto-spacer adjacent motif (PAM) and guiding RNAs (Jinek et al., 2012). Novel Cas9 proteins we discovered include some relatively short Type II-A and Type II-C Cas9 homologs (1,078-1,174 AA) that have potential for efficient virus-based packaging and delivery (Fig. 5). Furthermore, we determined corresponding putative *trans*-activating crRNAs (tracrRNAs) for Type II-A systems (Supplementary Fig. 12, panel B), which is instrumental in designing wild type crRNA:tracrRNA guides and synthetic single guide RNAs for Cas9 (Jinek et al., 2012). We further characterized the key elements of Type II systems for *L. jensenii*, *L. buchneri* and *L. mali* (Fig. 5), revealing the sequence diversity and structure conservation for the guide RNAs and their corresponding PAMs.

Phage and plasmid sequences were detected in 92% and 41% of the 213 genomes, respectively (Supplementary Fig. 13 and 14). Several synteny-based methods were used for predicting prophages, but the results were inconclusive and

subsequent manual analysis did little to improve this. Prediction of phage-specific genes was therefore used as an alternative and synteny-based methods of prophage prediction will be optimised for future studies. There is a trend towards an inverse correlation between abundance of CRISPR sequences and phage sequences that does not reach statistical significance (data now shown). Lactobacilli can have complex genome architecture (Raftis et al., 2014), and in many genomes multiple plasmids were detected (e.g. 6 plasmids predicted in both *L. parafarraginis* and *P. claussenii*; Supplementary Fig. 14). The phenomenon of very large plasmids exemplified by the sole genome sequence harbouring a megaplasmid in this analysis (the 380kb megaplasmid of *L. salivarius* DSM20555 (Felis et al., 2007)) substantially increases the number of plasmid-borne genes that are assigned to COGs for this genome (Supplementary Fig. 14). However, the influence of the megaplasmid on COG abundance is not evident on a genome-wide scale (Supplementary Fig. 15). These vectors open new avenues for genetic manipulation of model lactobacilli in the laboratory and for food-grade strain development. Furthermore, a diversity of insertion sequence (IS) elements was identified (Supplementary Fig. 16) including widespread IS families (IS3 is nearly universal), as well as sequences that selectively occur in particular niches (e.g. IS91 in dairy *L. casei* and *L. paracasei tolerans* and IS481 in brewing *L. paracollinoides*, *L. farraginis* and *P. inopinatus*). Altogether, mobile genetic elements and their occurrence reflect both the open pan-genome of lactobacilli and evolution by gene acquisition, and genome simplification and decay. Functionally, we also show that detected CRISPR spacer sequences can perfectly match target phage and plasmid sequences (Fig. 5), which is consistent with sequence-specific targeting of viruses by CRISPR-Cas adaptive systems. The findings from analysis of these 213 genomes corroborates previous reports implicating CRISPR-Cas systems in adaptive immunity against bacteriophages and plasmids in lactic acid bacteria used as starter cultures in food fermentation.

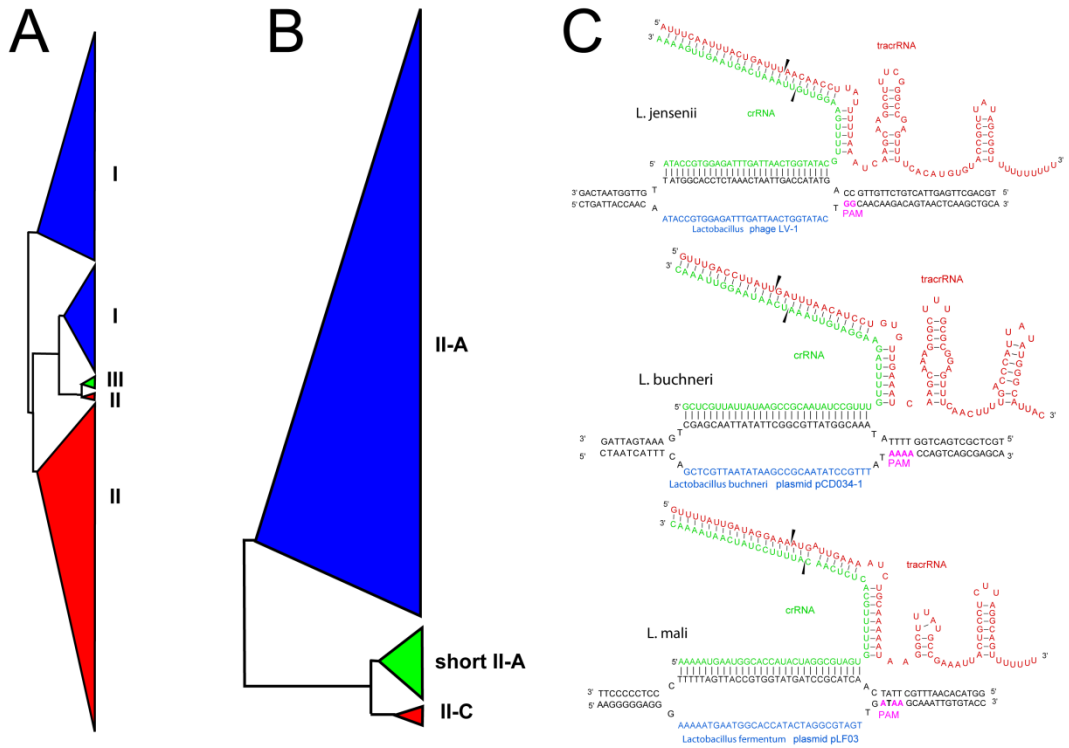


Figure 5: Comparative analysis of CRISPR sequences. The tree in panel A is derived from an alignment of the sequence of the universal Cas protein, Cas1, to create a phylogenetic tree based on the relatedness of all CRISPR-Cas systems in lactobacilli and closely related organisms. Types I, II and III are represented in blue, red and green, respectively. The tree in panel B is derived from an alignment of Cas9, the signature protein for Type II systems, to create a phylogenetic tree showing the relatedness of Cas9 proteins from Type II-A and II-C systems identified in lactobacilli and closely related organisms. A subset of short Type II-A Cas9 proteins is highlighted. In panel C, key guide sequences driving DNA targeting by Cas9 are shown for *L. jensenii*, *L. buchneri* and *L. mali*. Predicted crRNA, and tracrRNA sequences are shown at the top (red). Complementarity between CRISPR spacer sequences and target protospacer sequences (blue) in target nucleic acids is shown for phages and plasmids. The predicted protospacer-adjacent motif (PAM) sequences flanking the 3' end of the protospacer sequence are shown in green.

4 DISCUSSION

This *Lactobacillus* genome sequencing initiative provides genomic clarity for a genus bedevilled by phenotypic confusion and inconsistent phylogeny. We generated a resource dataset whose analysis explained the phenotypic diversity of lactobacilli and associated genera, and suggested new units for classification. The 200 genomes sequenced were from organisms spanning 9 genera and 174 species; including available *Oenococcus* and *Leuconostoc* genomes brought this to 11 genera and 185 species. We sequenced the genomes of *L. crustorum*, *L. parabrevis*, *L. pobuzihii* and *L. selangorensis* twice, but from different culture collections, and their sequence identity validated the sequencing and analysis pipelines. We elected to produce genomes of High Quality Draft standard (Chain et al., 2009), which is suitable for mining all relevant phylogenetic and functional information, and allows easy custom finishing as desired for genome regions of interest or whole genomes. Of the 200 type strains sequenced, 179 were previously unavailable on NCBI, which allows an unprecedented degree of integration of *Lactobacillus* genomics into taxonomic discussions and decisions. Since we started the sequencing phase, an additional 29 lactobacilli or candidate lactobacilli have been published in the literature; the definition of core genes and robust phylogeny described here will make their addition to the phylogenome easy once their genomes are sequenced.

Uncertainty surrounding species assignment and grouping into larger taxonomical units is undesirable, and it presents a considerable challenge for some bacteria such as those we termed here “the *Lactobacillus* Genus Complex”. Formal re-classification is the prerogative of systematic committees, but we examined phylogenomic approaches that might guide such classification. We first examined the most recent phylogeny (Salvetti et al., 2012) containing 16 phylogroups, and determined the frequency distribution of branch distances within phylogroup co-members and non-members (Supplementary Fig. 17, panel A1) based on the core gene tree (Fig. 2). We also calculated the frequency distribution of whole genome-wide genetic distance that is measured by the 1- TNI value (Supplementary Fig. 17, panel B1). The ideal phylogrouping that would yield non-intersecting curves was clearly not achieved through measurement of branch lengths or TNI values.

Therefore, we manually edited phylogroup membership primarily to concord with monophyletic clades, as well as to minimize the intersection area between curves (Supplementary Fig. 18). Although the TNI value distribution was still not discriminatory after optimizing the phylogroups (Supplementary Fig. 17, panel B2), we achieved superior separation of branch length distribution (Supplementary Fig. 17, panel A2). However, a stringent cut-off value for judging whether two strains belong to the same phylogroup could not be achieved, which may be due to unequal clock rates or speciation rates throughout the tree (which will be hard to determine based on current strain information). Nevertheless, the revised phylogrouping based on core genome comparison presented here can serve as the basis for discussions of formal re-classification.

Mobile replicons including bacteriophages and plasmids are a prominent feature of this group of bacteria, and have historically attracted attention because of their ability to extend the phenotype of a strain, or in the case of phage, to lyze starter or adjunct cultures. The data in this genome resource extend the knowledge base for exploiting the *Lactobacillus* mobilome. There is also a proportional abundance of systems to modulate the movement of these replicons. Collectively, our data reveal the widespread occurrence of diverse CRISPR-Cas immune systems in the genomes of lactobacilli, including a plethora of novel Type II systems with diverse Cas9 sequences. Of particular interest is the identification of a variety of Cas9 proteins that can be used in combination with novel guide sequences and various associated targeting motifs for flexible DNA targeting and cleavage. We anticipate that these novel systems will open new biotechnological avenues for next-generation Cas9-mediated genome editing in eukaryotes and prokaryotes. The broad occurrence of diverse CRISPR-Cas immune systems in lactobacilli in general also provides enormous potential for strain genotyping and enhancing phage resistance in industrial strains.

The genomic analysis highlights the remarkable diversity of pili in lactic acid bacteria. This also suggests that the pilus biogenesis, assembly, and also function may differ quite considerably between strains. To date, there have been only a few reports describing pili in *Lactobacillus* species other than *L. rhamnosus*. The present data offer a useful basis for future functional studies of these potentially piliated species from an environmental and evolutionary perspective.

Our data indicate that the *Lactobacillus* ancestor was facultatively heterofermentative, and that selective gene loss events have fine-tuned glycolysis/hexose/pentose metabolism in clade-specific patterns, against the back-drop of generalized gene loss and genome decay that characterizes the evolution of the Lactobacillales (Makarova et al., 2006). The selective pressures other than in the dairy environment are not well understood. Further evolutionary analyses are expected to resolve the presence of exceptions we described within major groups (characterized by a different genetic background compared with that of the whole group).

Apart from a pattern driven by genome reduction in animal-associated strains, we did not identify evidence for strong association between the niches of particular species and their genomic content (Supplementary Info.) though it must be recognized that the recorded isolation source of any given species may not necessarily be where it evolved. The strongly divergent patterns already illuminated by the current dataset for genes involved in carbohydrate management, proteolysis, surface protein production and destruction of foreign DNA provide a rational framework for species selection, trait browsing, replicon design and process optimization in fermentation and bioprocessing applications.

Author contributions

H.Z., Y.C. and P.W.O'T. conceived and designed the study, coordinated the project and partly wrote the manuscript. M.C.R., P.R. and A.W. cultured strains and extracted gDNA for sequencing. T.R.K., W.M.deV., P.W.C., R.B., A.McC. and H.M.B.H. designed analyses, performed bioinformatic analyses and partly wrote the manuscript; W.L. and

Y.S. performed validation experiments; X.Y. and C.G. assembled and annotated the genomes; Z.S., W.Z., X.Y., C.G. and R.Y. performed bioinformatic analyses; J.R. and

F.P.D. performed bioinformatic analyses and partly wrote the manuscript; S.A., E.S., G.E.F., O.O'S., R.P.R., A.E.B., J.P., J.C.C. and T.F.K. analysed the data and partly wrote the manuscript. I.B.J. supervised and performed bioinformatic analyses.

5 BIBLIOGRAPHY

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.,* 215**,** 403-10.

ALVAREZ-SIEIRO, P., MARTIN, M. C., REDRUELLO, B., DEL RIO, B., LADERO, V., PALANSKI, B. A., KHOSLA, C., FERNANDEZ, M. & ALVAREZ, M. A. 2014. Generation of food-grade recombinant *Lactobacillus casei* delivering *Myxococcus xanthus* prolyl endopeptidase. *Appl. Microbiol. Biotechnol.,* 98**,** 6689-700.

ARDITA, C. S., MERCANTE, J. W., KWON, Y. M., LUO, L., CRAWFORD, M. E., POWELL, D. N., JONES, R. M. & NEISH, A. S. 2014. Epithelial adhesion mediated by pilin SpaC is required for *Lactobacillus rhamnosus* GG-induced cellular responses. *Appl. Environ. Microbiol.,* 80**,** 5068-77.

BADEL, S., BERNARDI, T. & MICHAUD, P. 2011. New perspectives for Lactobacilli exopolysaccharides. *Biotechnol. Adv.,* 29**,** 54-66.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., RICHARDS, M., BOYAVAL, P., MOINEAU, S., ROMERO, D. A. & HORVATH, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315**,** 1709-12.

BARRANGOU, R. & MARRAFFINI, L. A. 2014. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell,* 54**,** 234-44.

BEG, Q. K., KAPOOR, M., MAHAJAN, L. & HOONDAL, G. S. 2001. Microbial xylanases and their industrial applications: a review. *Appl. Microbiol. Biotechnol.,* 56**,** 326-38.

BERGMAN, M., DEL PRETE, G., VAN KOOYK, Y. & APPELMELK, B. 2006. *Helicobacter pylori* phase variation, immune modulation and gastric autoimmunity. *Nat. Rev. Microbiol.,* 4**,** 151-9.

BERMUDEZ-HUMARAN, L. G., AUBRY, C., MOTTA, J. P., DERAISON, C., STEIDLER, L., VERGNOLLE, N., CHATEL, J. M. & LANGELLA, P. 2013. Engineering lactococci and lactobacilli for human health. *Curr. Opin. Microbiol.,* 16**,** 278-83.

BERNARDEAU, M., GUGUEN, M. & VERNOUX, J. P. 2006. Beneficial lactobacilli in food and feed: long-term use, biodiversity and proposals for specific and realistic safety assessments. *FEMS Microbiol. Rev.,* 30**,** 487-513.

BERNARDEAU, M., VERNOUX, J. P., HENRI-DUBERNET, S. & GUEGUEN, M. 2008. Safety assessment of dairy microorganisms: the Lactobacillus genus. *Int. J. Food Microbiol.,* 126**,** 278-85.

CANCHAYA, C., CLAESSON, M. J., FITZGERALD, G. F., VAN SINDEREN, D. & O'TOOLE, P. W. 2006. Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. *Microbiology,* 152**,** 3185-3196.

CASTILLO MARTINEZ, F. A., BALCIUNAS, E. M., SALGADO, J. M., DOMINGUEZ DOMINGUEZ, J. M., CONVERTI, A. & DE SOUZA OLIVEIRA, R. P. 2013. Lactic acid properties, production and applications: a review. *Trends Food Sci. Tech.,* 30**,** 70-83.

CHAIN, P. S., GRAFHAM, D. V., FULTON, R. S., FITZGERALD, M. G., HOSTETLER, J., MUZNY, D., ALI, J., BIRREN, B., BRUCE, D. C., BUHAY, C., COLE, J. R., DING, Y., DUGAN, S., FIELD, D., GARRITY, G. M., GIBBS, R., GRAVES, T., HAN, C. S., HARRISON, S. H., HIGHLANDER, S., HUGENHOLTZ, P., KHOURI, H. M., KODIRA, C. D., KOLKER, E., KYRPIDES, N. C., LANG, D., LAPIDUS, A., MALFATTI, S. A., MARKOWITZ, V., METHA, T., NELSON, K. E., PARKHILL, J., PITLUCK, S., QIN, X., READ, T. D., SCHMUTZ, J., SOZHAMANNAN, S., STERK, P., STRAUSBERG, R. L., SUTTON, G., THOMSON, N. R.,

TIEDJE, J. M., WEINSTOCK, G., WOLLAM, A. & DETTER, J. C. 2009. Genomics. Genome project standards in a new era of sequencing. *Science,* 326**,** 236-7.

CHAN, J. Z., HALACHEV, M. R., LOMAN, N. J., CONSTANTINIDOU, C. & PALLEN, M. J. 2012. Defining bacterial species in the genomic era: insights from the genus Acinetobacter. *BMC Microbiol,* 12**,** 302.

CHEN, J., YANG, X., CHEN, J., CEN, Z., GUO, C., JIN, T., YANG, R. & CUI, Y. 2013. *SISP: a Fast Species Identification System for Prokaryotes Based on Total Nucleotide Identity of Whole Genome Sequence* [Online]. Available: http://figshare.com/articles/SISP_a_Fast_Species_Identification_System_for_Prokaryotes_Based_on_Total_Nucleotide_Identity_of_Whole_Genome_Sequence/781226.

CHUNG, S. Y. & SUBBIAH, S. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure,* 4**,** 1123-7.

CHYLINSKI, K., MAKAROVA, K. S., CHARPENTIER, E. & KOONIN, E. V. 2014. Classification and evolution of type II CRISPR-Cas systems. *Nucleic. Acids Res.,* 42**,** 6091-105.

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE P, W. 2008. *Lactobacillus* phylogenomics - towards a reclassification of the genus. *Int. J. Syst. Evol. Microbiol.,* 58**,** 2945-2954.

COLLINS, M. D., RODRIGUES, U., ASH, C., AGUIRRE, M., FARROW, J. A. E., MARTINEZMURCIA, A., PHILLIPS, B. A., WILLIAMS, A. M. & WALLBANKS, S. 1991. Phylogenetic analysis of the genus *Lactobacillus* and related lactic-acid bacteria as determined by reverse-transcriptase sequencing of 16s ribosomal-RNA. *FEMS Microbiol. Letts.,* 77**,** 5-12.

DE JONG, A., VAN HEEL, A. J., KOK, J. & KUIPERS, O. P. 2010. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.,* 38**,** W647-51.

DELCHER, A. L., BRATKE, K. A., POWERS, E. C. & SALZBERG, S. L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics,* 23**,** 673-9.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.,* 32**,** 1792-7.

FELIS, G. E., MOLENAAR, D., DELLAGLIO, F. & VAN HYLCKAMA VLIEG, J. E. 2007. Dichotomy in post-genomic microbiology. *Nat. Biotechnol.,* 25**,** 848-9.

FOSTER, J. M., DAVIS, P. J., RAVERDY, S., SIBLEY, M. H., RALEIGH, E. A., KUMAR, S. & CARLOW, C. K. 2010. Evolution of bacterial phosphoglycerate mutases: non-homologous isofunctional enzymes undergoing gene losses, gains and lateral transfers. *PLoS One,* 5**,** e13576.

FRANCESCHINI, A., SZKLARCZYK, D., FRANKILD, S., KUHN, M., SIMONOVIC, M., ROTH, A., LIN, J., MINGUEZ, P., BORK, P., VON MERING, C. & JENSEN, L. J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res,* 41**,** D808-15.

FUJITA, K., OURA, F., NAGAMINE, N., KATAYAMA, T., HIRATAKE, J., SAKATA, K., KUMAGAI, H. & YAMAMOTO, K. 2005. Identification and molecular cloning of a novel glycoside hydrolase family of core 1 type O-glycan-specific endo-alpha-N-acetylgalactosaminidase from *Bifidobacterium longum*. *J. Biol. Chem.,* 280**,** 37415-22.

GORIS, J., KONSTANTINIDIS, K. T., KLAPPENBACH, J. A., COENYE, T., VANDAMME, P. & TIEDJE, J. M. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.,* 57**,** 81-91.

GRISSA, I., VERGNAUD, G. & POURCEL, C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics,* 8**,** 172.

GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.,* 52**,** 696-704.

HAFT, D. H., SELENGUT, J. D. & WHITE, O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res,* 31**,** 371-3.

HAMMES, W. P. & VOGEL, R. F. 1995. The genus *Lactobacillus*. *In:* WOOD, B. J. B. & HOLZAPFEL, W. H. (eds.) *The genera of Lactic Acid Bacteria.* Glasgow: Blackie Academic and Professional, UK.

HILL, C., GUARNER, F., REID, G., GIBSON, G. R., MERENSTEIN, D. J., POT, B., MORELLI, L., CANANI, R. B., FLINT, H. J., SALMINEN, S., CALDER, P. C. & SANDERS, M. E. 2014. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.*

JIANG, W., BIKARD, D., COX, D., ZHANG, F. & MARRAFFINI, L. A. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.,* 31**,** 233-9.

JINEK, M., CHYLINSKI, K., FONFARA, I., HAUER, M., DOUDNA, J. A. & CHARPENTIER, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science,* 337**,** 816-21.

KAGAWA, T. F., O'CONNELL, M. R., MOUAT, P., PAOLI, M., O'TOOLE, P. W. & COONEY, J. C. 2009. Model for substrate interactions in C5a peptidase from Streptococcus pyogenes: A 1.9 A crystal structure of the active form of ScpA. *J. Mol. Biol.,* 386**,** 754-72.

KANDLER, O. 1983. Carbohydrate metabolism in lactic acid bacteria. *Antonie Van Leeuwenhoek,* 49**,** 209-24.

KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res,* 42**,** D199-205.

KANKAINEN, M., PAULIN, L., TYNKKYNEN, S., VON OSSOWSKI, I., REUNANEN, J., PARTANEN, P., SATOKARI, R., VESTERLUND, S., HENDRICKX, A. P., LEBEER, S., DE KEERSMAECKER, S. C., VANDERLEYDEN, J., HAMALAINEN, T., LAUKKANEN, S., SALOVUORI, N., RITARI, J., ALATALO, E., KORPELA, R., MATTILA-SANDHOLM, T., LASSIG, A., HATAKKA, K., KINNUNEN, K. T., KARJALAINEN, H., SAXELIN, M., LAAKSO, K., SURAKKA, A., PALVA, A., SALUSJARVI, T., AUVINEN, P. & DE VOS, W. M. 2009. Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. *Proc Natl Acad Sci U S A,* 106**,** 17193-8.

KANT, R., BLOM, J., PALVA, A., SIEZEN, R. J. & DE VOS, W. M. 2011. Comparative genomics of *Lactobacillus*. *Microb. Biotechnol.,* 4**,** 323-32.

KATAYAMA, T., SAKUMA, A., KIMURA, T., MAKIMURA, Y., HIRATAKE, J., SAKATA, K., YAMANOI, T., KUMAGAI, H. & YAMAMOTO, K. 2004. Molecular cloning and characterization of *Bifidobacterium bifidum* 1,2-alpha-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95). *J. Bacteriol.,* 186**,** 4885-93.

KATOH, K. & TOH, H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.,* 9**,** 286-98.

KLAENHAMMER, T. R., KLEEREBEZEM, M., KOPP, M. V. & RESCIGNO, M. 2012. The impact of probiotics and prebiotics on the immune system. *Nat. Rev. Immunol.,* 12**,** 728-34.

LEISNER, J. J., VOGENSEN, F. K., KOLLMANN, J., AIDEH, B., VANDAMME, P., VANCANNEYT, M. & INGMER, H. 2008. alpha-Chitinase activity among lactic acid bacteria. *Syst Appl Microbiol,* 31**,** 151-6.

MAKAROVA, K., SLESAREV, A., WOLF, Y., SOROKIN, A., MIRKIN, B., KOONIN, E., PAVLOV, A., PAVLOVA, N., KARAMYCHEV, V., POLOUCHINE, N., SHAKHOVA, V., GRIGORIEV, I., LOU, Y., ROHKSAR, D., LUCAS, S., HUANG, K., GOODSTEIN, D. M., HAWKINS, T., PLENGVIDHYA, V., WELKER, D., HUGHES, J., GOH, Y., BENSON, A., BALDWIN, K., LEE, J. H., DIAZ-MUNIZ, I., DOSTI, B., SMEIANOV, V., WECHTER, W., BARABOTE, R., LORCA, G., ALTERMANN, E., BARRANGOU, R., GANESAN, B., XIE, Y., RAWSTHORNE,

H., TAMIR, D., PARKER, C., BREIDT, F., BROADBENT, J., HUTKINS, R., O'SULLIVAN, D., STEELE, J., UNLU, G., SAIER, M., KLAENHAMMER, T., RICHARDSON, P., KOZYAVKIN, S., WEIMER, B. & MILLS, D. 2006. Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci. U S A,* 103**,** 15611-6.

MARCO, M. L., DE VRIES, M. C., WELS, M., MOLENAAR, D., MANGELL, P., AHRNE, S., DE VOS, W. M., VAUGHAN, E. E. & KLEEREBEZEM, M. 2010. Convergence in probiotic *Lactobacillus* gut-adaptive responses in humans and mice. *ISME J.,* 4**,** 1481-4.

MOHAMADZADEH, M., DUONG, T., SANDWICK, S. J., HOOVER, T. & KLAENHAMMER, T. R. 2009. Dendritic cell targeting of *Bacillus anthracis* protective antigen expressed by *Lactobacillus acidophilus* protects mice from lethal challenge. *Proc. Natl. Acad. Sci. U S A,* 106**,** 4331-6.

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A. C. & KANEHISA, M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.,* 35**,** W182-5.

NAVARRE, W. W. & SCHNEEWIND, O. 1999. Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol Mol Biol Rev,* 63**,** 174-229.

OH, J. H. & VAN PIJKEREN, J. P. 2014. CRISPR-Cas9-assisted recombineering in Lactobacillus reuteri. *Nucleic Acids Res.,* 42**,** e131.

PLYUSNIN, I., HOLM, L. & KANKAINEN, M. 2009. LOCP--locating pilus operons in gram-positive bacteria. *Bioinformatics,* 25**,** 1187-8.

QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., HARTE, N., MULDER, N., APWEILER, R. & LOPEZ, R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res,* 33**,** W116-20.

RAFTIS, E. J., FORDE, B. M., CLAESSON, M. J. & O'TOOLE, P. W. 2014. Unusual genome complexity in Lactobacillus salivarius JCM1046. *BMC Genomics,* 15**,** 771.

REDDY, G., ALTAF, M., NAVEENA, B. J., VENKATESHWAR, M. & KUMAR, E. V. 2008. Amylolytic bacterial lactic acid fermentation - a review. *Biotechnol. Adv.,* 26**,** 22-34.

RODRÍGUEZ JIMÉNEZ, E. 2009. Dextranase in sugar industry: A review. *Sugar Tech.,* 11**,** 124-134.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-5.

SALVETTI, E., TORRIANI, S. & FELIS, G. E. 2012. The genus *Lactobacillus*: a taxonomic update. *Probiotics Antimic. Proteins,* 4**,** 217-226.

SHALLOM, D., GOLAN, G., SHOHAM, G. & SHOHAM, Y. 2004. Effect of dimer dissociation on activity and thermostability of the alpha-glucuronidase from *Geobacillus stearothermophilus*: dissecting the different oligomeric forms of family 67 glycoside hydrolases. *J. Bacteriol.,* 186**,** 6928-37.

SIEZEN, R. J. 1999. Multi-domain, cell-envelope proteinases of lactic acid bacteria. *Antonie Van Leeuwenhoek,* 76**,** 139-55.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. & NATALE, D. A. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4**,** 41.

TEAM, R. C. 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

VON OSSOWSKI, I., REUNANEN, J., SATOKARI, R., VESTERLUND, S., KANKAINEN, M., HUHTINEN, H., TYNKKYNEN, S., SALMINEN, S., DE VOS, W. M. & PALVA, A. 2010. Mucosal adhesion properties of the probiotic *Lactobacillus rhamnosus* GG SpaCBA and SpaFED pilin subunits. *Appl. Environ. Microbiol.,* 76**,** 2049-57.

VON SCHILLDE, M. A., HORMANNSPERGER, G., WEIHER, M., ALPERT, C. A., HAHNE, H., BAUERL, C., VAN HUYNEGEM, K., STEIDLER, L., HRNCIR, T., PEREZ-MARTINEZ, G., KUSTER, B. & HALLER, D. 2012. Lactocepin secreted by Lactobacillus exerts anti-inflammatory effects by selectively degrading proinflammatory chemokines. *Cell Host Microbe,* 11**,** 387-96.

WANG, L. & WISE, M. J. 2011. Glycogen with short average chain length enhances bacterial durability. *Naturwissenschaften,* 98**,** 719-29.

WU, D., JOSPIN, G. & EISEN, J. A. 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One,* 8**,** e77033.

WU, M. & SCOTT, A. J. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics,* 28**,** 1033-4.

ZHANG, Z. G., YE, Z. Q., YU, L. & SHI, P. 2011. Phylogenomic reconstruction of lactic acid bacteria: an update. *BMC Evol. Biol.,* 11**,** 1.

Chapter III

Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal

Published as:

Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal.

Harris HMB, Bourin MJB, Claesson MJ, O'Toole PW.

Microb Genom. 2017 Jun 13; 3(8): e000115.

doi: 10.1099/mgen.0.000115.

eCollection 2017 Aug.

PMID: 29026656

TABLE OF CONTENTS

1. INTRODUCTION 115
2. METHODS 117
2.1 Sequencing, assembly and annotation 117
2.2 Core-gene and single-gene phylogeny 118
2.3 Core-genome and pan-genome curves 119
2.4 Whole-genome comparisons: ANI and POCP 119
2.5 Additional methods sections 120
3. RESULTS AND DISCUSSION 121
3.1 A dataset of 42 genomes is sufficient to capture the *L. salivarius* core genome but not to capture the diversity of accessory genes 121
3.2 The core-gene phylogenetic tree of *L. salivarius* has similar topology to ANI whole genome clusters and single-gene phylogenies 123
3.3 Plasmids contribute considerably to *L. salivarius* genomic diversity 126
3.4 LPXTG-motif surface proteins are more numerous in strains harbouring multiple sortases and a putative pilus operon 129
3.5 The gene distributions of glycosyl hydrolases and glycosyl transferases show considerable evidence of gene loss and HGT 131
3.6 Host adaptation and gene conservation in EPS gene clusters 134
3.7 Bacteriocin gene content ranges from ubiquitous to strain-specific 137
4. CONCLUSIONS 139
5. DATA BIBLIOGRAPHY 140
6. BIBLIOGRAPHY 143

1 INTRODUCTION

The genus *Lactobacillus* is a diverse, paraphyletic group with a combined species and subspecies count of over 200 (Sun et al., 2015). Lactobacilli are Gram-positive, rod-shaped, non-spore-forming bacteria that inhabit a wide range of niches from soil and plants to the gastrointestinal tracts of humans and animals (Salvetti et al., 2012, Slover and Danziger, 2008). They are the largest group within the lactic acid bacteria (LAB) and one of the most important bacterial groups involved in food microbiology and human nutrition because of their fermentative and probiotic properties (Salvetti et al., 2012).

Several pivotal studies have called for a reclassification of the *Lactobacillus* genus (Claesson et al., 2008, Salvetti et al., 2012, Sun et al., 2015) while others have provided detailed characterisation of its diversity (Salvetti et al., 2012, Claesson et al., 2008, Sun et al., 2015, Zheng et al., 2015a, Canchaya et al., 2006, Kant et al., 2011). Sun *et al* recently conducted an international genome sequencing initiative of the lactobacilli that revealed that the genus was more diverse than a typical taxonomic family and that confirmed that *Leuconostoc*, *Oenococcus*, *Weissella*, *Pediococcus* and *Fructobacillus* all branch from within the *Lactobacillus* phylogenetic tree (Sun et al., 2015).

Numerous studies have also focused on the comparative genomics of individual *Lactobacillus* species, highlighting considerable intraspecific genomic diversity among strains (Forde et al., 2011, Broadbent et al., 2012, Cremonesi et al., 2012, Douillard et al., 2013, Smokvina et al., 2013, Ojala et al., 2014, Senan et al., 2014, MM et al., 2015, Wegmann et al., 2015, Zheng et al., 2015b, Raftis et al., 2011, Martino et al., 2016). One species that has been repeatedly isolated from the gastro-intestinal tracts of humans and animals and that has potential probiotic properties is the facultatively heterofermentative species, *Lactobacillus salivarius* (Claesson et al., 2006, Messaoudi et al., 2013, Neville and O'Toole, 2010).

The genome of *L. salivarius* UCC118 was first characterised by Claesson *et al* and shown to have a multi-replicon organisation with a single *repA*-type megaplasmid and two smaller plasmids. The megaplasmid harboured genes with an array of functions including bile salt hydrolysis, carbohydrate metabolism and genes

that complete the pentose phosphate pathway. The study concluded that the megaplasmid increased the metabolic flexibility and competitiveness of the species (Claesson et al., 2006). A previous study also identified a novel bacteriocin, Abp118, encoded by the megaplasmid of UCC118 (Flynn et al., 2002). Two exopolysaccharide (EPS) production gene clusters were found on the UCC118 chromosome, which share homology and synteny with other *L. salivarius* strains (Raftis et al., 2011). EPS, among other bacterial factors, has been implicated in bile tolerance in species including *L. rhamnosus* (Koskenniemi et al., 2011).

Two studies showed that other strains of *L. salivarius* share a similar multi-replicon organisation to that of UCC118, each having a homologous *repA*-type megaplasmid and a varying number of smaller plasmids from none to two (Li et al., 2007, Fang et al., 2008). Several strains have more complicated architectures: JCM1046, JCM1047 and AH43348 all have a linear megaplasmid (Li et al., 2007) as well as a *repA*-type megaplasmid while JCM1046 also has an additional circular megaplasmid (Raftis et al., 2014). The varying presence of plasmids in *L. salivarius* as well as the variation in size of the megaplasmids (Li et al., 2007) (100-380 kb) suggests that there is considerable functional diversity across the strains. This variation is not limited to the plasmids. Raftis *et al* used the two chromosomal EPS clusters of UCC118 as a reference in a comparative genome hybridisation (CGH) experiment that revealed considerable divergence in gene synteny and gene presence among 33 strains of *L. salivarius* (Raftis et al., 2011).

The previous study by Raftis *et al* constituted a largely non-bioinformatic analysis of *L. salivarius* strains but nevertheless revealed interesting functional differences (Raftis et al., 2011). The present study seeks to conduct a fully bioinformatic analysis of the phylogeny and functional divergence in an expanded dataset of 42 *L. salivarius* genomes. The constraint of using a reference strain (UCC118) that CGH demands is not a limiting factor of the present study, and strain-specific as well as clade-specific genes and functions can be identified by comparative genomics that would otherwise be excluded. We focussed on the analysis of numerous functional traits and we also provide an overall whole-genome view of the relatedness of the strains and the extent of their diversity.

2 METHODS

2.1 SEQUENCING, ASSEMBLY AND ANNOTATION

The genomes of a panel of 29 *L. salivarius* strains were sequenced by Macrogen Ltd. (Beotkkot-ro-Geumcheon-qu, Seoul, Rep. of Korea) using the HiSeq platform and 100 bp paired-end reads. This dataset was supplemented by 13 *L. salivarius* genomes (5 complete and 8 draft) that were available in NCBI databases. *L. hayakitensis* DSM18933 was also included in the study as a related out-group. The dataset included both genome sequences for the type strain from two different culture collections (DSM20555^T and ATCC11741^T) to test the robustness of the methods.

Reads for the 29 sequenced genomes were assembled using Velvet (v1.2.10) (Zerbino, 2010) with a kmer count of 61, and with expected coverage and coverage cut-off both set to 'auto', allowing Velvet to infer these values. Nucleotide coverages were all high (>100x) and assembly statistics are available in Table S1. Mauve (v2.4.0) (Rissman et al., 2009) was used to reorder and reorient draft contigs relative to the complete genome of UCC118. Additional quality checks are described in Supplementary methods.

Genes were predicted using three different gene prediction software: Glimmer3 (v3.02) (Delcher et al., 2007), GeneMark.HMM (v1.1) (Besemer et al., 2001) and MetaGene (Noguchi et al., 2006). In cases where software predictions disagreed on the correct start site for a gene, the longest predicted gene sequence was chosen. Genes predicted by one software only were still included in the dataset in order to minimise false negative gene predictions.

The issue of multi-copy genes such as the 16S rRNA gene is not addressed in this study. Our dataset contains a majority of draft genome sequences where assembly software often fails to assemble multiple copies of identical or almost identical genes due to ambiguous placement of reads. Similar genes that posed no problem for assembly software were included in gene counts analysis.

The amino acid sequences of predicted genes were BLASTed (blastp) against the Kyoto Encyclopaedia of Genes and Genomes database (KEGG) (Ogata et al.,

1999), the Clusters of Orthologous Groups database (COG) (Tatusov et al., 1997) and the non-redundant NCBI database (*www.ncbi.nlm.nih.gov)* to assign functional annotation. BLAST thresholds for assigning the function of a reference sequence to a query gene were 40% identity, 50% alignment length to the query gene and a BLAST bit score of 60. Prediction and annotation of specific functional groups in this study are described in Supplementary methods.

2.2 CORE-GENE AND SINGLE-GENE PHYLOGENY

QuartetS (Yu et al., 2011) was used to cluster predicted genes (amino acid sequences) into orthologs. It does this by calculating the reciprocal best BLAST hits (RBBs) between the genes of each pair of genomes and performing two-stage clustering (single linkage and Markov clustering) on the RBBs. BLAST thresholds were 40% identity, 50% alignment length of the query gene and a BLAST bit score of 50. For clustering the RBBs, an MCL inflation value of 3 and a minimum cluster size of 2 were used.

The 42 *L. salivarius* genomes and the *L. hayakitensis* DSM18933 genome combined had a predicted core genome of 938 genes. For each genome, these 938 genes were concatenated and the resulting sequences were aligned across the genome set using Muscle (v3.8.31) (Edgar, 2004). Gap regions were removed in R (v3.2.3) (R Core Team, 2015) where each amino acid position in the alignment is a column and all columns with at least one gap are excluded. RAxML (v8.0.22) (Stamatakis, 2014) was used to generate a bootstrapped tree (100 iterations) from the core gene alignment using a PROTCATCPREV model and FigTree (v1.4.0) (Morariu et al., 2009) was used to visualise the tree, which was rooted on *L. hayakitensis* DSM18933. The root branch was artificially shortened to provide greater visual discrimination across *L. salivarius* sub-clades so all other branches are informative relative to each other.

To supplement the core-gene phylogeny, 4 single-gene phylogenies were also generated based on nucleotide sequences using the above methods and a GTRCAT model. These 4 genes are *groEL*, *rpsB, parB* and *rpoA,* which were identified in each genome using reference sequences from UCC118.

2.3 CORE-GENOME AND PAN-GENOME CURVES

A binary gene matrix modified from the QuartetS output was used to generate core-, pan- and new-gene curves in R. *L. hayakitensis* DSM18933 was excluded from this analysis. Unique genes that were excluded by QuartetS (due to a minimum cluster size of 2) were also added to the matrix at this point. The number of core, pan and new genes were calculated by starting with two genomes and sequentially adding genomes, one at a time, until all 42 genomes were included. This procedure was repeated 1,000 times, each time the order of the matrix being permuted to randomise the order of addition of genomes. Median values along with the variation from each permutation were recorded and plotted using R. In order to assess the open or closed nature of a pan-genome, the $\log_{10}$ median values for the new-gene curve were also plotted where a slope of less than 1 is interpreted as belonging to an open pan-genome ($\alpha < 1$) (Tettelin et al., 2008). The R code for permuting the binary-gene matrix and creating a pan-genome matrix for plotting the pan-genome curve is on figshare (see Data Bibliography; data file 1). Similar code was used for the core- and new-gene curves (data file 2 and data file 3, respectively).

2.4 WHOLE-GENOME COMPARISONS: ANI AND POCP

Two whole-genome comparative metrics were used to supplement the core-gene and single-gene phylogenies. Average Nucleotide Identity (ANI) (Goris et al., 2007) and Percentage of Conserved Proteins (POCP) (Qin et al., 2014) are two widely employed methods that seek to provide accurate species and genus cut-off values, respectively. To calculate ANI values for each pair of genomes, an ANI Perl script was downloaded (https://github.com/chjp/ANI/blob/master/ANI.pl) and implemented. Qin *et al* (Qin et al., 2014) did not provide a POCP script so an in-house script was written using the same formula and BLAST thresholds listed in their paper. The script used for POCP calculation is on figshare (see Data Bibliography; data file 4).

2.5 ADDITIONAL METHODS SECTIONS

Additional descriptions of Methods can be found in Supplementary Methods. These carry the sub-headings, ‘Quality assessment of genomes’, ‘Assigning contigs to replicons’ and ‘Specific functional groups’.

3 RESULTS AND DISCUSSION

3.1 A DATASET OF 42 GENOMES IS SUFFICIENT TO CAPTURE THE *L. SALIVARIUS* CORE GENOME BUT NOT TO CAPTURE THE DIVERSITY OF ACCESSORY GENES

The core genome of *L. salivarius* consisted of 1,236 genes. Applying a leave-one-out-strategy to the 42 *L. salivarius* genomes and re-computing the core genome shows that it varies from 1,236 to 1,246 with 1,281 as an outlier when JCM1230 is excluded. Table S1 shows that the JCM1230 strain sequenced in this study possesses no plasmids, which explains why the core genome increased so much when the strain was excluded - the absence of a megaplasmid excludes all extrachromosomal genes from being part of the core genome. Li *et al* (Li et al., 2007) identified a *repA*-type megaplasmid in JCM1230 and predicted its size to be approximately 100 kb. It is difficult to explain the absence of plasmid sequences in JCM1230 in the current study: the megaplasmid might have been artificially excluded by a procedural artefact during the DNA extraction/preparation procedure or, alternatively, since 100 kb is the smallest *repA*-type megaplasmid in the Li *et al* (Li et al., 2007) dataset, the strain may have lost the megaplasmid *in vitro* during laboratory passage.

Fig. 1(a) shows the core gene curve for the 42 *L. salivarius* genomes. The curve starts to plateau after the addition of only a few genomes and has substantially levelled out by genome number 42. This suggests that a dataset of 42 genomes is sufficient to define the core genome of *L. salivarius*. Hutchison *et al* (Hutchison et al., 2016) recently conducted a study on the synthesis of a minimal bacterial genome that required 473 genes to survive under lab conditions. Like many other species, the core genome size of *L. salivarius*, with approximately 1,200 genes, suggests that most of the core genes of a specific group of bacteria are necessary for processes outside of basic cell viability such as niche adaptation and interaction with competitors and pathogens.

The accessory genome of the 42 *L. salivarius* genomes (excluding unique genes) consists of 3,057 gene clusters ranging from 802 genes present in only two genomes to 109 genes present in 41 genomes (all but one). Fig. 1(b) shows the pan-

genome curve (core and accessory, including unique genes) for the 42 *L. salivarius* genomes. The steep slope indicates that the current dataset is not large enough to define the accessory genome of *L. salivarius* and that the addition of more genomes from other strains would continue to increase the size of the accessory gene set. Fig. 1(c) shows that the new-gene curve plateaus off at a steady addition of approximately 100 genes per genome. The new-gene curve is a combination of accessory homologous genes and strain-specific genes although homologs might still exist that are not RBBs or that fall below cut-off values.

Overall, the data presented in Fig. 1 supports the model for an open pan-genome (Fig. 1(d); $\alpha < 1$) (Tettelin et al., 2008) whereby an expanding dataset of *L. salivarius* genomes will continue to acquire novel genes. Variation in the presence of genes within species is brought about by two main processes, HGT and gene decay, both of which apparently began to act upon all *L. salivarius* strains after they diverged from their common ancestor, leading to the intra-specific variation observed in this dataset.

This intra-specific variation can be summarised in a very general sense using the median number of genes per replicon with the first and third quartiles representing inter-genome variation: chromosome = 1,737 (1,685, 1,844); megaplasmid = 249 (216, 283) and small plasmid = 47.5 (23.5, 89.5).

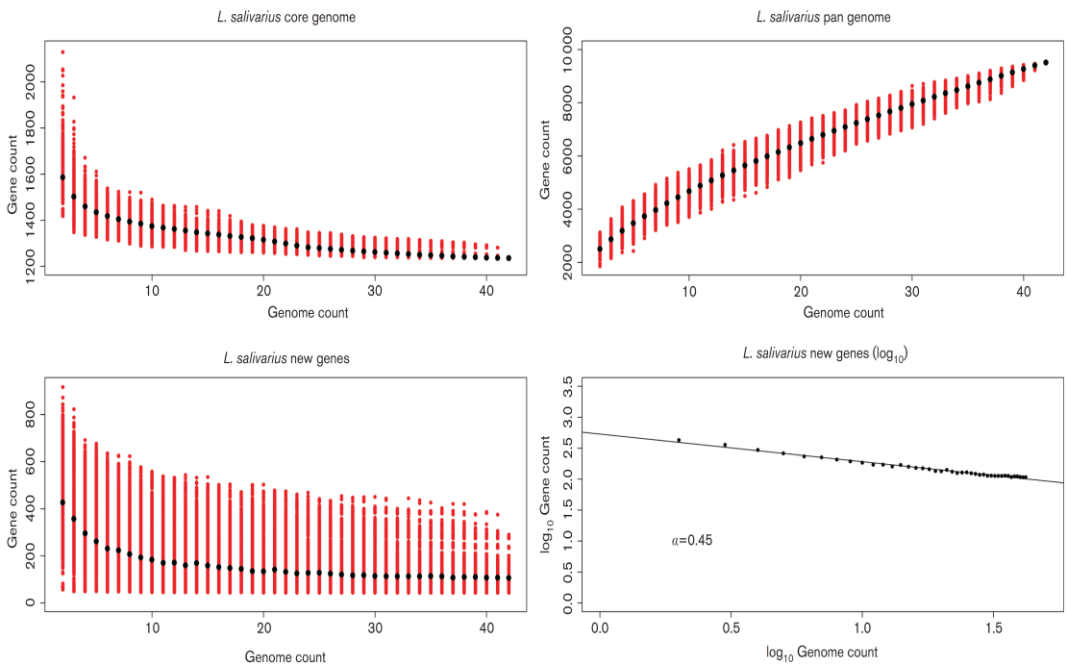


Fig. 1: A dataset of 42 genomes is not sufficient to define the *L. salivarius* pan-genome. The four panels show, with the sequential addition of 42 *L. salivarius* genomes (x-axis), the decrease in core genes (panel a; top-left), the increase in total genes (panel b; top-right), the decrease in new genes (panel c; bottom-left) and the log of the decrease in new genes (panel d; bottom-right). Genes are counted as orthologous gene families (% identity >= 40 and % alignment length >=50) except for genes unique to each genome. The order of addition of genomes has been permuted 1,000 times. Red dots show the variation in values while black dots show the median value. An alpha value of 0.44 shows that the pan genome of *L. salivarius* is open ($\alpha < 1$).

3.2 THE CORE-GENE PHYLOGENETIC TREE OF *L. SALIVARIUS* HAS SIMILAR TOPOLOGY TO ANI WHOLE GENOME CLUSTERS AND SINGLE-GENE PHYLOGENIES

Fig. 2 shows the core-gene phylogeny of *L. salivarius*, rooted on *L. hayakitensis* DSM18933. The bootstrap values are high, indicating a robust tree topology and the length of most of the branches leading to the nodes suggests that

some divergence has occurred even in more closely related strains. Note that the out-group branch (DSM18933) has been shortened for this analysis (see Methods), but the scale indicating 0.003 substitutions per amino acid position can still be applied to all *L. salivarius* branches. A few sub-clades have little to no outer branch lengths, reflecting a lack of phylogenetic divergence. LMG14476 and LMG14477 have a difference of only 8 SNPs in the predicted core of 938 genes even though they were isolated from different sources (Table S1). Three strains isolated from the oral cavity - gul1 and gul2 (isolated in the same study), and DSM20555^T (independent isolate) - also show limited phylogenetic divergence (8-19 SNPs). ATCC11741^T is the same *L. salivarius* type strain as DSM20555^T from another culture collection and they have a difference of 0 SNPs in the predicted core of 938 genes, highlighting the limited accrual of variation over short periods of time during vertical gene transfer. A similar case can be observed for three strains - AH4231, AH4331 and AH43348 (17-48 SNPs) - all isolated from the human ileocecal region in the same study and between UCC118 and AH43324 (54 SNPs) also isolated from the human ileocecal region. In contrast to these sub-clades, CCUG44481 (an animal isolate) and CCUG38008 (a human gall isolate) have the most divergent core genome across all 42 *L. salivarius* strains (3,643 SNPs).

Average Nucleotide Identity (ANI) (Goris et al., 2007) was also used to cluster *L. salivarius* strains. Fig. 3 shows a heatmap of ANI values where the clustering of strains is largely in agreement with the core-gene phylogeny of Fig. 2. *L. hayakitensis* was excluded from the heatmap so an unrooted clustering is presented. ANI was designed as a method to identify whether a particular strain belongs within a species, using a cut-off value of 95% as the species boundary (Goris et al., 2007). In terms of its use of homologous sequences, ANI can be compared with the core-gene phylogenetic method, although it uses nucleotide sequences and includes homologous inter-genic regions. Discrepancies between the two tree topologies are likely due to differences in computing similarity scores from intra-genic amino acid sequences and intra/inter-genic nucleotide sequences. The lowest ANI value across the *L. salivarius* strains is 96.8% between JCM1047 (isolated from swine intestine) and CECT5713 (isolated from human breast milk), indicating that all strains belong to the same species.

Single-gene phylogenies were also constructed using 4 marker genes - *groEL*, *rpsB*, *parB* and *rpoA*. When sub-clades had sufficient phylogenetic signal,

bootstrap values were high and agreed with the tree topology of the core-gene phylogeny in Fig. 2. On average, however, the phylogenetic signal of the trees was too low to make reliable comparisons, reflecting the limits of building single-gene trees to study the evolutionary history within a species, especially since gene sequences had to be aligned at the nucleotide level to see what little divergence there was across strains for these genes. The tree for *parB* is included as Fig. S1 since it shows the most phylogenetic signal of the 4 genes.

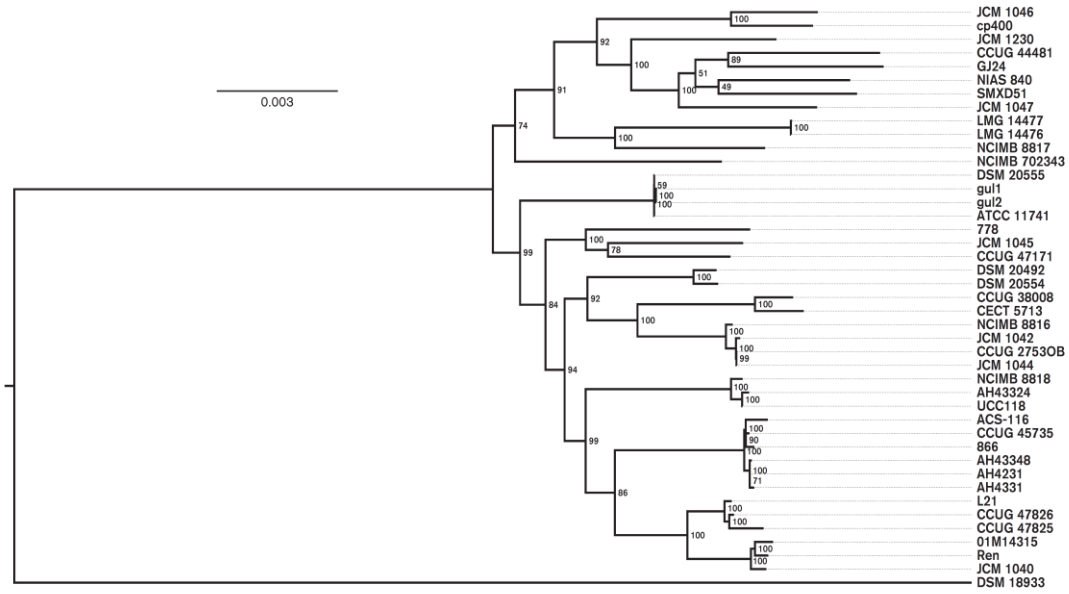


Fig. 2: A phylogenetic tree generated from 938 core genes shows considerable variation in divergence across strains. Branch lengths (solid black lines) represent evolutionary divergence and strain labels are lined up for ease of comparison (dashed lines). Bootstrap values are included to show robustness of tree topology. The tree is rooted on *L. hayakitensis* DSM18933 and this branch is artificially reduced to provide a clearer visualisation of the other branch lengths relative to each other. The scale bar shows average number of amino acid substitutions per site.

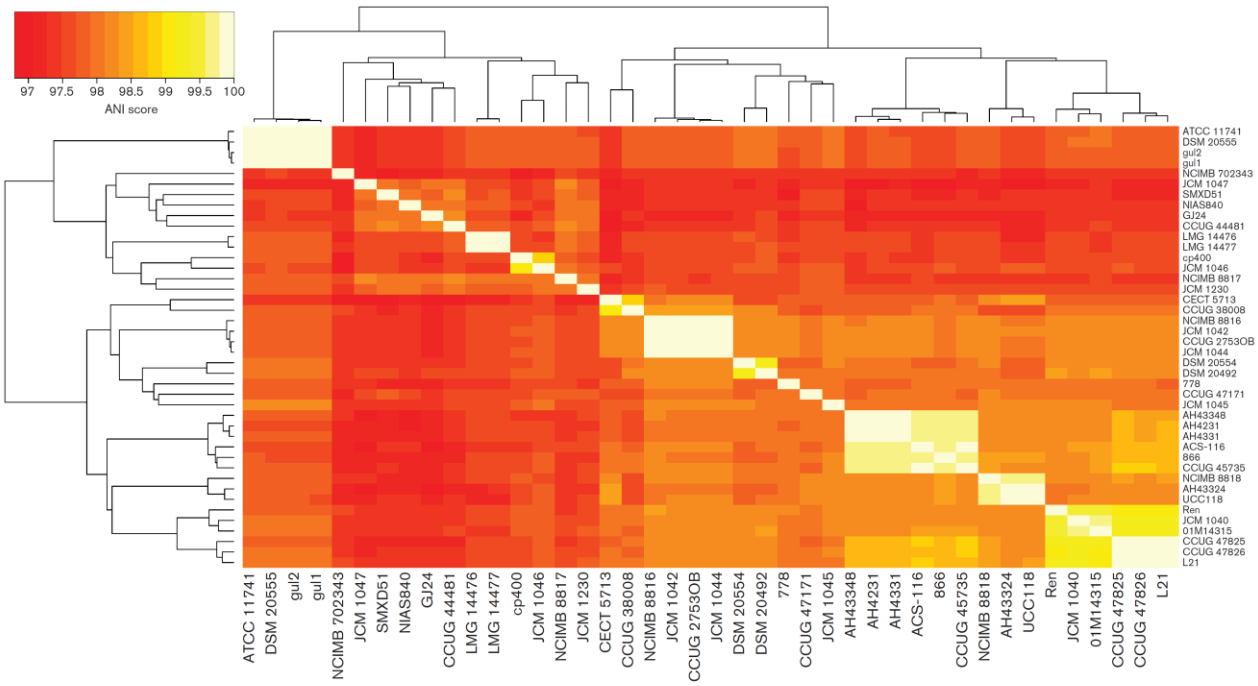


Fig. 3: **Clustering of pair-wise average nucleotide identity (ANI) scores agrees largely with the clustering of the core-gene tree in Fig. 2.** The colour key (top-left) shows a gradation of colour from red to orange to yellow to white representing increasing genome-genome similarity. Euclidean distance and complete linkage clustering were used to cluster rows and columns. *L. hayakitensis* DSM18933 is excluded.

3.3 PLASMIDS CONTRIBUTE CONSIDERABLY TO *L. SALIVARIUS* GENOMIC DIVERSITY

Li *et al* have already shown that there is considerable size variation in *L. salivarius repA*-type megaplasmids ranging from 100 kb (JCM1230) to 380 kb (DSM20555^{T}) (Li et al., 2007). This suggests that there is comparable variation in functional diversity due to the high coding density of prokaryotic replicons. The number of predicted genes on the *repA*-type megaplasmids that we predicted ranged from 165 genes in NIAS840 to 408 genes in cp400. NIAS840 has a complete genome sequence while that of strain cp400 is a draft, suggesting that closed genomes are not a factor for bias when predicting the number of genes on

megaplasmids. The lack of plasmids in *L. hayakitensis* DSM18933 was not discussed when the strain was published (Morita et al., 2007) and plasmid absence has no effect on the conclusion that the *repA*-type megaplasmid was acquired early in *L. salivarius* evolution (Li et al., 2007). The possible technical reasons for the loss of a megaplasmid in JCM1230 have been covered in a previous section. Table S2 shows the BLAST results of three *repA*-type marker genes - *repA*, *repE* and *parA* - against the contigs of each genome. If contigs were assigned to replicons accurately, it is expected that BLAST hits for each gene would lie on predicted *repA*-type megaplasmid contigs. This is indeed the case with all three genes having between 93-100% identities over their full length aligned to a *repA*-type megaplasmid contig, usually all three genes aligning to the same contig. Exceptions include JCM1230, which had no BLAST hits due to its missing megaplasmid, AH43348, which had an extra *parA* gene on a predicted *repA*-type megaplasmid contig and *L. hayakitensis* DSM18933, which has a *repA* gene and a *parA* gene on a predicted chromosomal contig. The *repA* and *parA* genes of DSM18933 have a lower identity than the other hits (79% and 87%, respectively) and it is possible that these genes belong to an unidentified megaplasmid, although there was no mention of extrachromosomal sequences in the original species/strain description (Morita et al., 2007).

Several strains in the dataset also possess linear megaplasmids that have little homology to the *repA*-type megaplasmid, a finding that was first documented in Li *et al* (Li *et al*., 2007). These strains are JCM1046, JCM1047 and AH43348. The linear megaplasmids of JCM1046 and JCM1047 show high sequence similarity: two predicted contigs in the draft genome of JCM1047 cover most of the complete linear megaplasmid of JCM1046 (pLMP1046) with a high percentage identity. The genome of AH43348 is a draft made up of 114 contigs so the linear megaplasmid could only be predicted by sequence homology with other linear megaplasmids from the database of *Lactobacillus* NCBI plasmids (see Supplementary methods). The contigs of AH43348 had very little homology to pLMP1046; however, several contigs do cover most of a second megaplasmid present in NIAS840 aside from the contigs that align to *repA*-type megaplasmids. The second megaplasmid of NIAS840 was not described as being circular or linear (Ham et al., 2011) and it is possible that this megaplasmid is actually homologous to the linear megaplasmid of AH43348. An alternative explanation is that both AH43348 and NIAS840 have two circular megaplasmids; this would mean that the homology-based method used in this study

failed to predict the linear megaplasmid of AH43348, instead assigning its genes to the chromosome. SMXD51 is predicted to have an additional large plasmid as well as a *repA*-type megaplasmid; its draft genome is made up of 10 contigs, 6 belonging to the chromosome and the remaining 4 described as a 143 kb megaplasmid, an 85 kb large plasmid and two small plasmids (31 kb and 9 kb) (Kergourlay et al., 2012). We found that the 143 kb and the 85 kb plasmids both align over most of their sequence to different regions of the *repA*-type megaplasmid of UCC118 (pMP118), together adding up to over 94% of its length. This suggests that these two sequences are not separate plasmids, but together make up the *repA*-type megaplasmid of SMXD51 - a finding made more probable by the fact that the available SMXD51 genome is a draft genome.

The smaller plasmids show even greater variation. Our findings (Table S1) suggest that 15 strains have no small plasmids, 20 strains have a single small plasmid and 8 strains have two small plasmids. The number of predicted genes on the small plasmids ranged from 11 in a GJ24 plasmid to 144 in an AH4231 plasmid. Many of these plasmids show high-level homology to the two endogenous plasmids described by Fang *et al* in UCC118 (Fang et al., 2008). The small plasmid of JCM1046 (pCTN1046) is quite distinct from those in UCC118 and shares homology with a plasmid in SMXD51, a relationship first described in Raftis *et al* (Raftis et al., 2014).

Fig. 4 shows a general summary of functional diversity across the replicons for each strain using COG categories. The absence of megaplasmids in DSM18933 and JCM1230 is evident along with the absence of smaller plasmids in 15 strains. The proportional allocation of genes to COGs shows much more similarity across chromosomal genes than across those on the megaplasmids or the plasmids, reflecting the accessory nature of extrachromosomal DNA. The proportions (and raw counts) of genes involved in translation and ribosomal structure is much higher on the chromosomes, reflecting the complexity of chromosomal cellular machinery related to protein production when compared to that of the plasmids. All three replicon groups have a large number of genes with unknown function, highlighting current limits to annotation, but also the need for greater experimental investigation. The mobilome gene category is much higher as a percentage in the plasmids; this makes sense due to the different selection pressures acting on plasmids and it can be speculated that it benefits prophages and transposases to use the higher copy number and conjugative ability of plasmids to multiply.

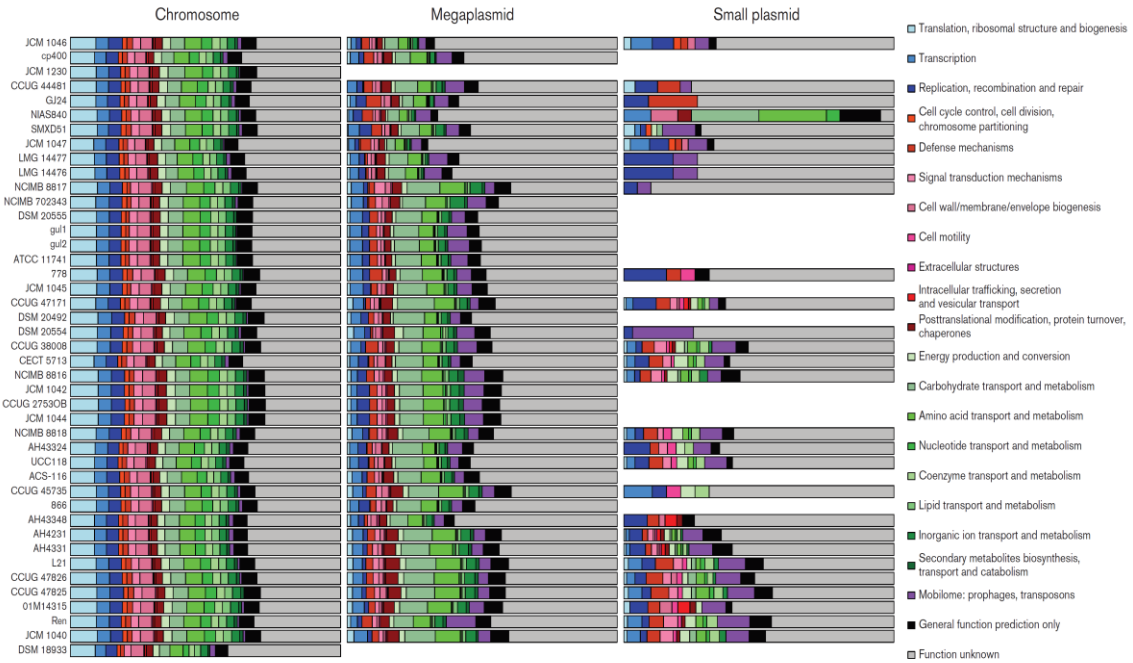


Fig. 4: **Proportion of genes assigned to each major COG category shows considerable variation across strains and plasmids.** Colours and order of COG categories in each bar from left to right match the colour legend from top to bottom. The order of the strains (bars) reflects the order of the core-gene tree in Fig. 2. Genes are separated into chromosomal, megaplasmid and plasmid genes. In cases where genomes have multiple plasmids or megaplasmids, the COG counts were combined. Note that genes assigned to the linear megaplasmids of AH43348, JCM1046 and JCM1047 are also included in the barplots for megaplasmids. The absence of plasmids from a particular genome is represented by the absence of a bar for that category.

3.4 LPXTG-MOTIF SURFACE PROTEINS ARE MORE NUMEROUS IN STRAINS HARBOURING MULTIPLE SORTASES AND A PUTATIVE PILUS OPERON

Sortases are important enzymes for recognising and anchoring surface proteins containing an LPXTG motif, and sortase-anchored surface proteins are often involved in the interaction of a bacterium with its surrounding environment (Call and Klaenhammer, 2013). In *L. salivarius*, this includes host-bacterium interactions since

most strains have been isolated from human or animal sources. Fig. 5 shows the gene counts for sortases, pilus genes and genes with an LPXTG motif.

All 43 genomes have at least one sortase A gene, the "housekeeping" sortase, that typically acts on many protein targets and is considered to be essential for the survival of most Gram-positive bacteria. Additionally, 7 genomes have an extra sortase A and 5 of these have a sortase C gene. All 5 strains with a sortase C have a putative pilus operon, confirming previous studies that describe the role of sortase C in pilus construction (Spirig et al., 2011). The extra sortase A in strains with a pilus operon suggests that this gene is a more specific sortase A with some role in the formation of pili. However, two strains, NIAS840 and AH43348, also have an additional sortase A gene, but they lack a pilus operon. We described in a previous section that the non-*repA*-type megaplasmid (presumably linear based on Li *et al* (Li et al., 2007)) of AH43348 has a strong homology to the second megaplasmid in NIAS840. The extra sortase A gene in these two strains lies on this extra megaplasmid (speculatively linear) and it presumably acts on gene products with an LPXTG motif encoded by this replicon. Four of the 5 strains with pilus operons belong to the DSM20555^{T} sub-clade (4 genomes) where 3 are isolated from the oral cavity and ATCC11741^{T} is a reference strain from the Human Microbiome Project (http://www.hmpdacc.org). Pili are commonly involved in adhesion and their production in this sub-clade might reflect an adaptation to the oral environment by allowing the bacterial cell to adhere to the tooth surface or underlying dentine. JCM1047 is a swine intestinal isolate and it is not clear why it is the only other strain with a predicted pilus operon, except that the presence of pili surely has an adaptive role in the intestine as well as the oral cavity.

The range of values for gene products with an LPXTG motif is partly explained by the number of sortase genes and the presence of pilus operons, with more genes being present in strains with multiple sortases and a pilus operon. *L. hayakitensis* DSM18933 has the most genes containing an LPXTG motif (n=18). This suggests that there might have been selective pressure leading to a reduced number of cell-surface and secreted proteins with an LPXTG motif in *L. salivarius*.

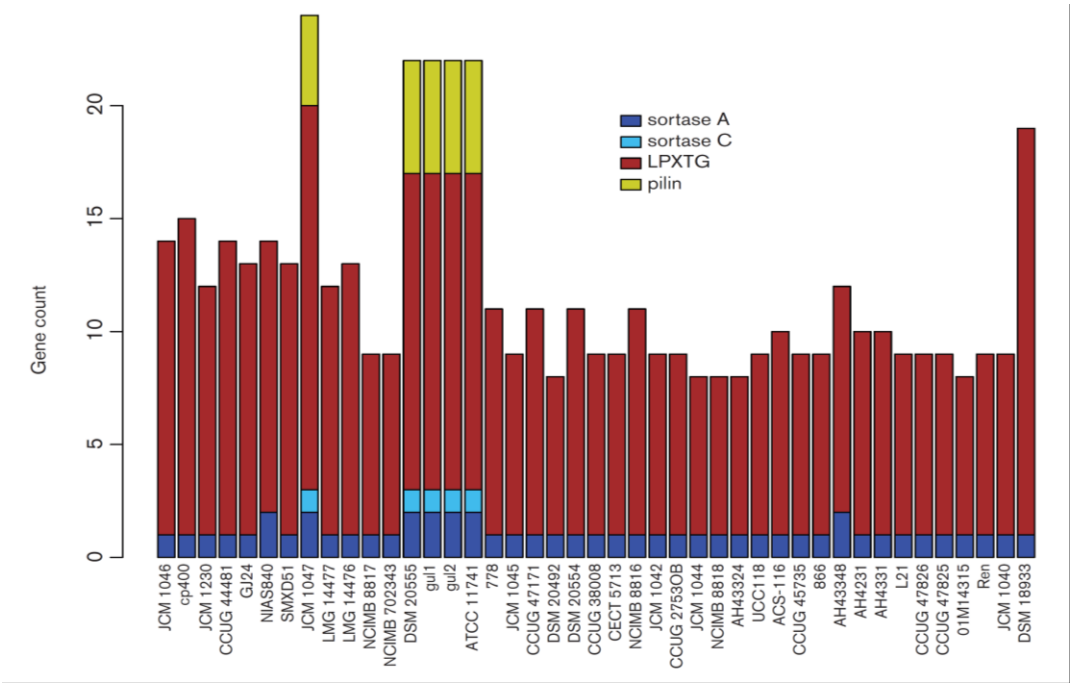


Fig. 5: Gene counts for sortase families, LPXTG motifs and potential pilus clusters are all positively correlated. Genes are assigned to 4 categories - sortase A, sortase C, LPXTG and pilin – and coloured according to the in-figure legend. The order of strains (bars) from left to right reflects the order of the core-gene tree from top to bottom in Fig. 2.

3.5 THE GENE DISTRIBUTIONS OF GLYCOSYL HYDROLASES AND GLYCOSYL TRANSFERASES SHOW CONSIDERABLE EVIDENCE OF GENE LOSS AND HGT

GHs and GTs are two large and important groups of genes that are responsible for the hydrolysis (or modification) and synthesis, respectively, of the glycosidic bonds of carbohydrates. Fig. 6 and Fig. 7 show the distribution and abundance of genes according to their GH and GT families across the 42 *L. salivarius* strains and *L. hayakitensis* DSM18933, separated into their respective replicons.

There is no correlation between the number of GHs and the number of GTs per strain in this dataset (Spearman rho = -0.07; p = 0.67), showing the independence of a strain's ability to synthesise carbohydrates compared to its ability to break them

down. This is not surprising since the selective pressures acting on genes that break down particular carbohydrates are largely determined by the availability of that substrate in the environment while carbohydrate synthesis can lead to complex interactive traits such as EPS, which vary in structure, composition and function depending on the biotic and abiotic environmental factors and the species of bacteria in question (Ciszek-Lenda, 2011).

For both GHs and GTs, the majority of genes reside on the chromosome (GH = 808/900; GT = 1,313/1,322), but there is considerably more extrachromosomal diversity for GHs than GTs and no GTs are located on the smaller plasmids. These results indicate that GHs are horizontally acquired more frequently than GTs in *L. salivarius*. GT families also appear to be more stable on the chromosome compared to GHs with 10 out of 13 GT families being present in 39 strains or greater while GHs have only 7 out of 17. Greater retention of GT genes across the dataset suggests that the relevant functions of carbohydrate synthesis are under greater selective pressure across all strains, whereas GH gene retention is more variable due to the dynamic and changeable nature of carbohydrate availability in typical environments for *L. salivarius* cells.

Numerous gene families for both GHs and GTs are present in all 43 genomes and found on the chromosomes only. For GHs, these are GH13, GH32 and GH73; for GTs, these are GT26, GT28, GT41 and GT51. All these families have numerous predicted substrates and functional properties and their absence from extra chromosomal replicons suggests that these genes are important for cell processes independent of particular niches. More interesting are the families that are present in all 42 *L. salivarius* genomes but absent from *L. hayakitensis* DSM18933 or, alternatively, absent from all 42 *L. salivarius* but present in DSM18933. These families are GH2 and GT32 (present in *L. salivarius* only), and GH68 (present in *L. hayakitensis* only). GH68 is a levansucrase and present in DSM18933 only while GH2 and GT32 are quite general and act on multiple substrates. Levansucrase enzymes, unlike sucrases, are localised almost entirely extracellularly and they contribute to 60% of extracellular sucrase activity (Goncalves, 2015). The presence of levansucrase in DSM18933 suggests that this strain is more adapted to the breakdown of sucrose – an ability that may compensate for the fact that this strain has the lowest number of GH genes (n=12) in this dataset and the lowest number of GH families (n=9) along with 01M14315, DSM20492 and SMXD51.

A few other GH and GT families have a very narrow distribution. GH70, a dextransucrase, is limited to CCUG44481 and GJ24 – a branch pair isolated from different sources. A gene for GH119, an α-amylase, is found only on the *repA*-type megaplasmid of JCM1046. Peptidoglycan lyase – an enzyme that can hydrolyse the cell walls of bacteria - is found on the smaller plasmids of JCM1046 and JCM1047, both isolates from the swine intestine. GT27 and GT92 are limited to the chromosomes of 5 strains: the sub-clade of 4 strains containing DSM20555^{T} and the singleton, NCIMB8817.

The distribution of genes across the strains in these two major functional groups shows considerable gene loss and HGT with very limited association of GH and GT families with isolation source.

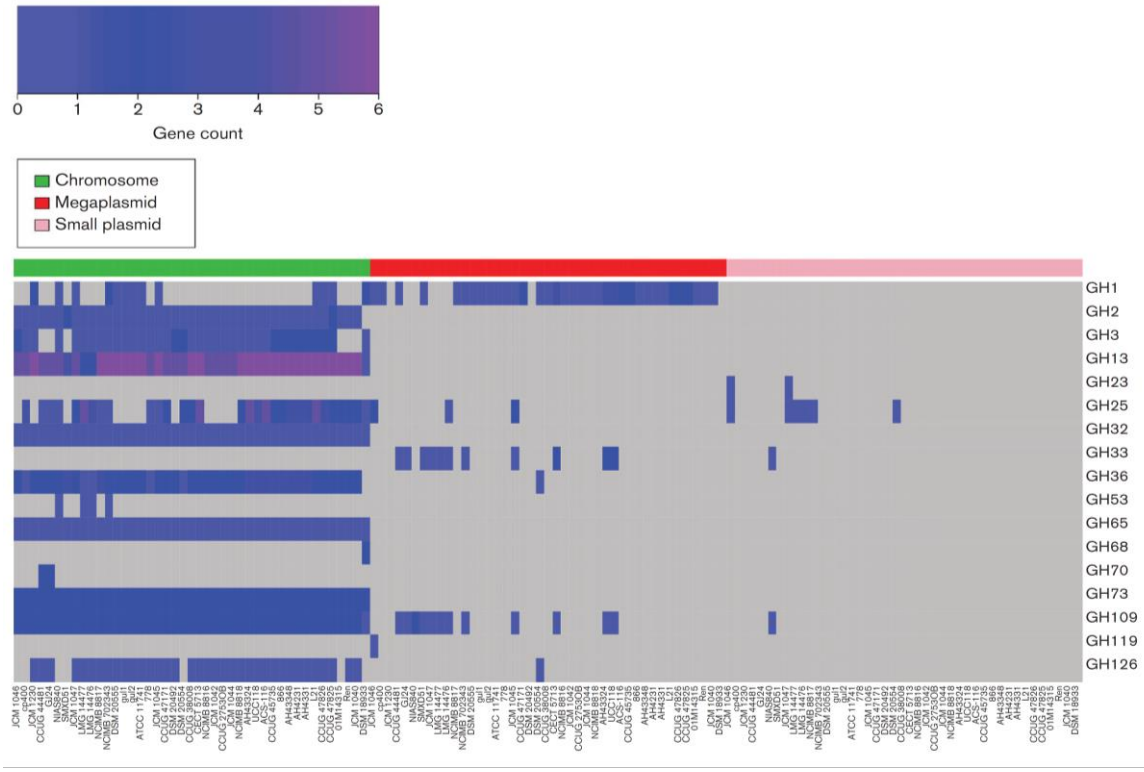


Fig. 6: Gene counts for GH families suggest regular HGT and gene loss events. The colour key (top-left) shows a gradation of colour from blue to dark blue to purple as the gene count increases. Grey represents a gene count of zero. GH genes are separated into chromosomal, megaplasmid and plasmid genes. For each replicon group, the order of strains (columns) from left to right reflects the order of the core-gene tree from top to bottom in Fig. 2.

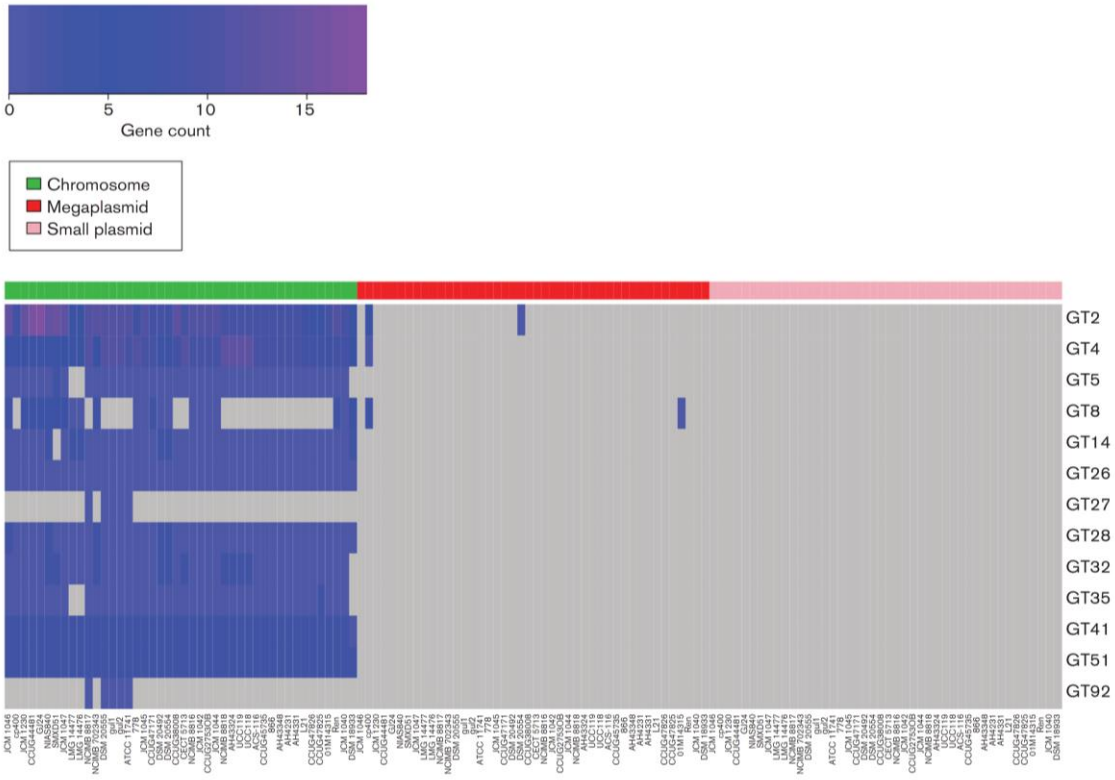


Fig. 7: Gene counts for GT families suggest more restricted HGT than that which occurs for GHs. The colour key (top-left) shows a gradation of colour from blue to dark blue to purple as the gene count increases. Grey represents a gene count of zero. GT genes are separated into chromosomal, megaplasmid and plasmid genes. For each replicon group, the order of strains (columns) from left to right reflects the order of the core-gene tree from top to bottom in Fig. 2.

3.6 HOST ADAPTATION AND GENE CONSERVATION IN EPS GENE CLUSTERS

L. salivarius UCC118 EPS cluster 1 is located on the chromosome and is composed of 21 genes spread across 23 kb. Twenty-nine *L. salivarius* strains harbour at least 18 genes from UCC118 EPS cluster 1 and the other 13 strains do not have the cluster in their genomes (Fig. 8). Interestingly, the presence of EPS cluster 1 is correlated with the core-gene tree (Fig. 2). The majority of strains in the top sub-clade from JCM1046 to NCIMB702343 lack EPS cluster 1. Two other strains, DSM20492 and DSM20554, are located in the middle of the tree and do not harbour the cluster either. DSM18933 lacks EPS cluster 1, suggesting that either the common

ancestor of *L. salivarius* acquired the cluster through HGT after the split from *L. hayakitenis* or, alternatively, that DSM18933 lost the cluster through gene decay.

Another interesting point is that 9 of the 13 strains lacking EPS cluster 1 were isolated from animal samples and only 3 were isolated from human samples (one strain does not have a known origin). In contrast to this, the majority of strains harbouring EPS cluster 1 have a human origin, suggesting that EPS cluster 1 is not essential for the survival of *L. salivarius* as a species, but it might code for an adaptive trait to the human GIT.

L. salivarius UCC118 EPS cluster 2 is also located on the chromosome and is composed of 28 genes spread across 33 kb. The two physical extremities of the EPS cluster 2 are shared by all the strains (Fig. 9; from LSL_1574 to LSL_1569 and from LSL_1551 to LSL_1547). However, variations exist in the middle of EPS cluster 2 and 6 groups were identified as described in Fig. S2. Group 1 contained strains harbouring all the UCC118 EPS cluster 2 genes while group 6 had only the 2 extremities of the cluster.

The central part of the cluster varies in the *L. salivarius* strains compared to the reference strain, UCC118. This region contained the majority of glycosyltransferases and EPS biosynthesis-related proteins in UCC118 EPS cluster 2. Glycosyltransferases are involved in the addition of sugar subunits to the growing EPS chain. A difference in the glycosyltransferase composition suggests potential variation in EPS structure. These results show that the organisation of EPS cluster 2 is not conserved in most *L. salivarius* strains. Indeed, only 4 strains belong to group 1: UCC118, AH43324, CECT5713 and NCIMB8818. Interestingly, potential probiotic activities have been described for CECT5713 (Perez-Cano et al., 2010) and UCC118 (Flynn et al., 2002).

EPS produced by strains of lactobacilli are suspected to play a role in the strain's probiotic activity (Lebeer et al., 2008). *L. salivarius* heteropolysaccharide production is controlled by EPS clusters and the structure of *Lactobacillus* EPS clusters has been described as highly conserved (Patten and Laws, 2015), although discussion in this area is still very much open - a fact that is highlighted in *L. salivarius* EPS clusters that vary considerably in both their gene synteny and in the presence of particular genes.

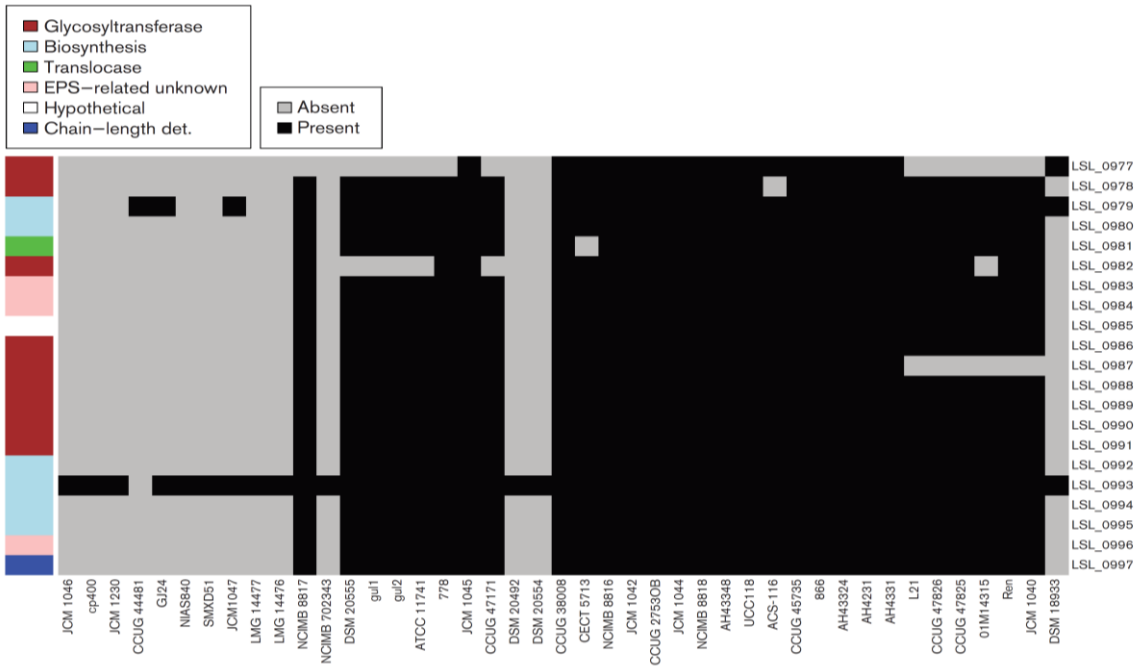


Fig. 8: EPS cluster 1 is absent from some strains of *L. salivarius*. Grey represents gene absence and black represents gene presence. The genes (rows) are ordered according to synteny in UCC118, which is used as a reference. The order of strains (columns) from left to right reflects the order of the core-gene tree from top to bottom in Fig. 2. The colour legend defines genes within the cluster in terms of general function.

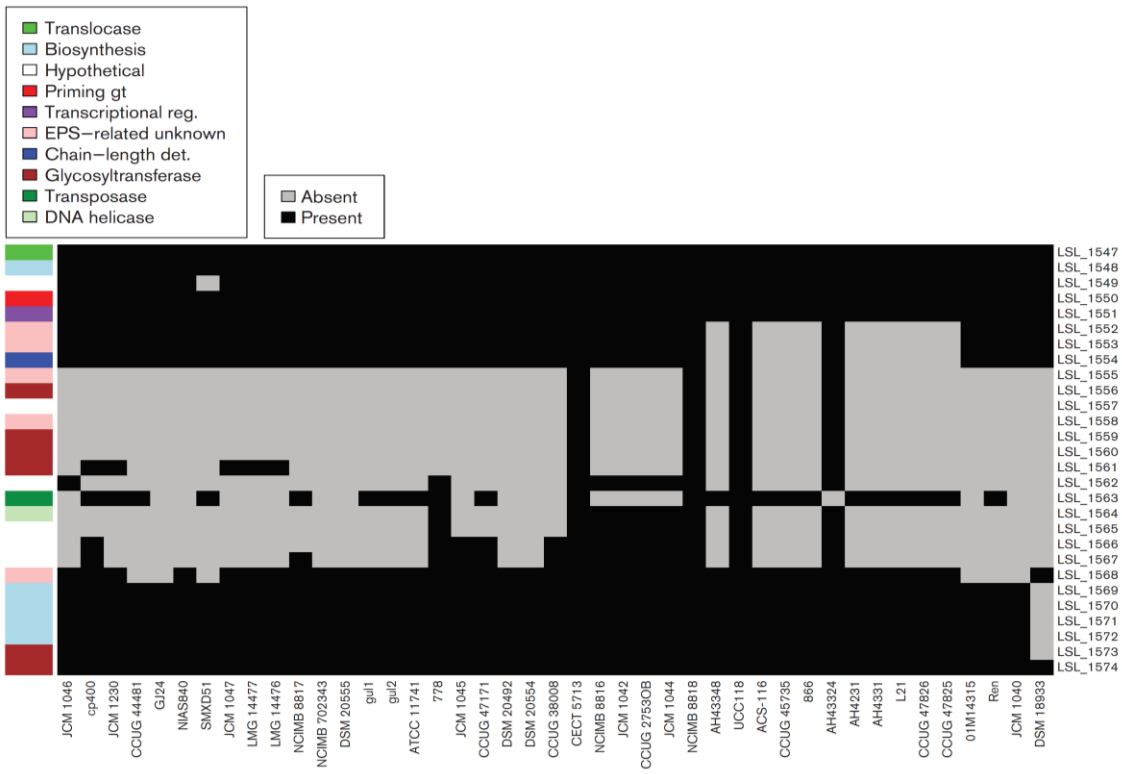


Fig. 9: EPS cluster 2 shows variable presence of genes at the centre in *L. salivarius*. Grey represents gene absence and black represents gene presence. The genes (rows) are ordered according to synteny in UCC118, which is used as a reference. The order of strains (columns) from left to right reflects the order of the core-gene tree from top to bottom in Fig. 2. The colour legend defines genes within the cluster in terms of general function.

3.7 BACTERIOCIN GENE CONTENT RANGES FROM UBIQUITOUS TO STRAIN-SPECIFIC

Flynn *et al* identified a small, heat-stable bacteriocin, Abp118, in UCC118 that showed considerable antimicrobial activity (Flynn et al., 2002). This bacteriocin is identified as salivaricin P by Bagel3, which has close homology to Abp118 since they differ by only two amino acids (Barrett et al., 2007). Homologs of Abp118 along with their surrounding genes (Areas of Interest; AOIs) are present in 22 strains of *L. salivarius* in this study (Table S3). In all 22 cases, this bacteriocin is found on the *repA*-type circular megaplasmid and appears to have no strong association with a

particular isolation source, but its distribution on the core-gene tree (Fig. 2) is associated with several sub-clades including the UCC118 branch (n=3), the AH43348 branch (n=6) and a few small sub-branches (of n=2) and singletons. It is interesting that some of the strains lack this bacteriocin; the size and functional variation of the *repA*-type megaplasmid highlights the fast evolutionary rate that these replicons undergo, perhaps losing Abp118 if bacteria co-inhabiting the same environment did not compete strongly with *L. salivarius* for limiting resources.

A number of other bacteriocins are also present in the *L. salivarius* strains in this dataset. All 43 strains possess between 1 and 4 enterolysin genes. The N-terminals of these bacteriocins have considerable sequence homology to a bacteriophage lysin and they act to degrade the bacterial cell-wall in a range of genera including enterococci, pediococci, lactococci and lactobacilli (Nilsen et al., 2003). LS2, an extremely heat- and pH-stable peptide with anti-listerial activity (Busarcevic and Dalgalarrondo, 2012)7, is confined to the NCIMB8816 sub-clade (n=4) and shows homology to bacteriocins in several oral streptococci. The two-strain sub-clade consisting of CCUG44481 and GJ24 is the only branch to harbour a plantaricin S while MR10B is present on the small plasmid of three strains – JCM1046, JCM1047 and DSM20554. A cluster of three bacteriocins is present on two divergent strains, CCUG44481 and CCUG47171, harbouring plantaricin NC8, lactacin F and acidocin LF221B. The distribution of bacteriocins in this dataset gives an indication of HGT: LS2 is confined to a single sub-clade and was likely transferred into the megaplasmid of the ancestor of these 4 strains; MR10B is present on the only small plasmid in 3 divergent strains.

The production of bacteriocins gives a strain an obvious competitive advantage since it inhibits similar strains and species that may compete strongly for limiting resources. Specific environments impose different biotic and abiotic factors and the details of microbial competition and horizontal transfer of genes (including bacteriocin genes) are dependent on a complicated interplay among these factors, potentially explaining the scattered distribution of bacteriocin genes in this dataset.

4 CONCLUSIONS

We conducted a comparative genomic study of 42 strains of *L. salivarius* and a closely related out-group, *L. hayakitenis* DSM18933. Previous comparative studies show that there is considerable functional and phylogenetic diversity across *Lactobacillus* species. Smaller scale intra-specific studies focusing on single *Lactobacillus* species highlight the continuation of this trend across strains.

We demonstrate that *L. salivarius* has an open pan-genome and that all major functional groups described show considerable functional variation across strains, often displaying greater similarity within sub-clusters as opposed to niche-specific trends. Variation in gene function is greater across the megaplasmids than across the chromosomes and greater across the smaller plasmids than across the megaplasmids. The level of functional variation revealed in *L. salivarius* suggests that strain-specific properties can potentially be applied to commercial areas of human health and nutrition such as probiotics and food preservation.

5 DATA BIBLIOGRAPHY

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163248; GenBank accession: MSCR00000000 (01M14315)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163249; GenBank accession: NBEY00000000 (AH4231)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163250; GenBank accession: NBEX00000000 (AH4331)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163251; GenBank accession: NBEW00000000 (AH43324)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163252; GenBank accession: NBEV00000000 (AH43348)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163253; GenBank accession: NBEU00000000 (CCuG2753OB)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163254; GenBank accession: NBET00000000 (CCuG38008)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163255; GenBank accession: NBES00000000 (CCuG44481)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163256; GenBank accession: NBER00000000 (CCuG45735)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163257; GenBank accession: NBEQ00000000 (CCuG47171)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163258; GenBank accession: NBEP00000000 (CCuG47825)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163259; GenBank accession: NBEO00000000 (CCuG47826)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163260; GenBank accession: NBEN00000000 (DSM20492)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163261; GenBank accession: NBEM00000000 (DSM20554)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163262; GenBank accession: NBEL00000000 (gul1)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163263; GenBank accession: NBEK00000000 (gul2)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163264; GenBank accession: NBEJ00000000 (JCM1040)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163265; GenBank accession: NBEI00000000 (JCM1042)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163266; GenBank accession: NBEH00000000 (JCM1044)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163267; GenBank accession: NBEG00000000 (JCM1045)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163268; GenBank accession: NBEF00000000 (JCM1047)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163269; GenBank accession: NBEE00000000 (JCM1230)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163270; GenBank accession: NBED00000000 (L21)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163271; GenBank accession: NBEC00000000 (LMG14476)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163272; GenBank accession: NBEB00000000 (LMG14477)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163273; GenBank accession: NBEA00000000 (NCIMB702343)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163274; GenBank accession: NBDZ00000000 (NCIMB8816)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163275; GenBank accession: NBDY00000000 (NCIMB8817)

Hugh Harris, Genbank; BioProject ID: PRJNA357984; BioSample accession: SAMN06163276; GenBank accession: NBDX00000000 (NCIMB8818)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577917.v1 (data file 1)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577947.v1 (data file 2)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577950.v1 (data file 3)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577953.v1 (data file 4)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577956.v1 (data file 5)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577965.v1 (data file 6)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577971.v1 (data file 7)

Hugh Harris, dx.doi.org/10.6084/m9.figshare.4577977.v1 (data file 8)

6 BIBLIOGRAPHY

BARRETT, E., HAYES, M., O'CONNOR, P., GARDINER, G., FITZGERALD, G. F., STANTON, C., ROSS, R. P. & HILL, C. 2007. Salivaricin P, one of a family of two-component antilisterial bacteriocins produced by intestinal isolates of Lactobacillus salivarius. *Appl Environ Microbiol,* 73**,** 3719-23.

BESEMER, J., LOMSADZE, A. & BORODOVSKY, M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res,* 29**,** 2607-18.

BROADBENT, J. R., NEENO-ECKWALL, E. C., STAHL, B., TANDEE, K., CAI, H., MOROVIC, W., HORVATH, P., HEIDENREICH, J., PERNA, N. T., BARRANGOU, R. & STEELE, J. L. 2012. Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. *BMC Genomics,* 13**,** 533.

BUSARCEVIC, M. & DALGALARRONDO, M. 2012. Purification and genetic characterisation of the novel bacteriocin LS2 produced by the human oral strain Lactobacillus salivarius BGHO1. *Int J Antimicrob Agents,* 40**,** 127-34.

CALL, E. K. & KLAENHAMMER, T. R. 2013. Relevance and application of sortase and sortase-dependent proteins in lactic acid bacteria. *Front Microbiol,* 4.

CANCHAYA, C., CLAESSON, M. J., FITZGERALD, G. F., VAN SINDEREN, D. & O'TOOLE, P. W. 2006. Diversity of the genus Lactobacillus revealed by comparative genomics of five species. *Microbiology,* 152**,** 3185-96.

CISZEK-LENDA, M. 2011. Review paper
 Biological functions of exopolysaccharides from probiotic bacteria. *Central European Journal of Immunology,* 36**,** 51-55.

CLAESSON, M. J., LI, Y., LEAHY, S., CANCHAYA, C., VAN PIJKEREN, J. P., CERDENO-TARRAGA, A. M., PARKHILL, J., FLYNN, S., O'SULLIVAN, G. C., COLLINS, J. K., HIGGINS, D., SHANAHAN, F., FITZGERALD, G. F., VAN SINDEREN, D. & O'TOOLE, P. W. 2006. Multireplicon genome architecture of Lactobacillus salivarius. *Proc Natl Acad Sci U S A,* 103**,** 6718-23.

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE, P. W. 2008. Lactobacillus phylogenomics--towards a reclassification of the genus. *Int J Syst Evol Microbiol,* 58**,** 2945-54.

CREMONESI, P., CHESSA, S. & CASTIGLIONI, B. 2012. Genome sequence and analysis of Lactobacillus helveticus. *Front Microbiol,* 3**,** 435.

DELCHER, A. L., BRATKE, K. A., POWERS, E. C. & SALZBERG, S. L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics,* 23**,** 673-9.

DOUILLARD, F. P., RIBBERA, A., KANT, R., PIETILA, T. E., JARVINEN, H. M., MESSING, M., RANDAZZO, C. L., PAULIN, L., LAINE, P., RITARI, J., CAGGIA, C., LAHTEINEN, T., BROUNS, S. J., SATOKARI, R., VON OSSOWSKI, I., REUNANEN, J., PALVA, A. & DE VOS, W. M. 2013. Comparative genomic and functional analysis of 100 Lactobacillus rhamnosus strains and their comparison with strain GG. *PLoS Genet,* 9**,** e1003683.

EDGAR, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

FANG, F., FLYNN, S., LI, Y., CLAESSON, M. J., VAN PIJKEREN, J. P., COLLINS, J. K., VAN SINDEREN, D. & O'TOOLE, P. W. 2008. Characterization of endogenous plasmids from Lactobacillus salivarius UCC118. *Appl Environ Microbiol,* 74**,** 3216-28.

FLYNN, S., VAN SINDEREN, D., THORNTON, G. M., HOLO, H., NES, I. F. & COLLINS, J. K. 2002. Characterization of the genetic locus responsible for the production of ABP-118, a

novel bacteriocin produced by the probiotic bacterium Lactobacillus salivarius subsp. salivarius UCC118. *Microbiology,* 148**,** 973-84.

FORDE, B. M., NEVILLE, B. A., O'DONNELL, M. M., RIBOULET-BISSON, E., CLAESSON, M. J., COGHLAN, A., ROSS, R. P. & O'TOOLE, P. W. 2011. Genome sequences and comparative genomics of two Lactobacillus ruminis strains from the bovine and human intestinal tracts. *Microb Cell Fact,* 10 Suppl 1**,** S13.

GONCALVES, B. C. M. 2015. *Levan And Levansucrase-A Mini Review*, International Journal of Scientific & Technology Research.

GORIS, J., KONSTANTINIDIS, K. T., KLAPPENBACH, J. A., COENYE, T., VANDAMME, P. & TIEDJE, J. M. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol,* 57**,** 81-91.

HAM, J. S., KIM, H. W., SEOL, K. H., JANG, A., JEONG, S. G., OH, M. H., KIM, D. H., KANG, D. K., KIM, G. B. & CHA, C. J. 2011. Genome sequence of Lactobacillus salivarius NIAS840, isolated from chicken intestine. *J Bacteriol,* 193**,** 5551-2.

HUTCHISON, C. A., CHUANG, R.-Y., NOSKOV, V. N., ASSAD-GARCIA, N., DEERINCK, T. J., ELLISMAN, M. H., GILL, J., KANNAN, K., KARAS, B. J., MA, L., PELLETIER, J. F., QI, Z.-Q., RICHTER, R. A., STRYCHALSKI, E. A., SUN, L., SUZUKI, Y., TSVETANOVA, B., WISE, K. S., SMITH, H. O., GLASS, J. I., MERRYMAN, C., GIBSON, D. G. & VENTER, J. C. 2016. Design and synthesis of a minimal bacterial genome. *Science,* 351.

KANDLER, O. 1983. Carbohydrate metabolism in lactic acid bacteria. *Antonie Van Leeuwenhoek,* 49**,** 209-24.

KANT, R., BLOM, J., PALVA, A., SIEZEN, R. J. & DE VOS, W. M. 2011. Comparative genomics of Lactobacillus. *Microb Biotechnol,* 4**,** 323-32.

KERGOURLAY, G., MESSAOUDI, S., DOUSSET, X. & PREVOST, H. 2012. Genome sequence of Lactobacillus salivarius SMXD51, a potential probiotic strain isolated from chicken cecum, showing anti-campylobacter activity. *J Bacteriol,* 194**,** 3008-9.

KOSKENNIEMI, K., LAAKSO, K., KOPONEN, J., KANKAINEN, M., GRECO, D., AUVINEN, P., SAVIJOKI, K., NYMAN, T. A., SURAKKA, A., SALUSJARVI, T., DE VOS, W. M., TYNKKYNEN, S., KALKKINEN, N. & VARMANEN, P. 2011. Proteomics and transcriptomics characterization of bile stress response in probiotic Lactobacillus rhamnosus GG. *Mol Cell Proteomics,* 10**,** M110 002741.

LEBEER, S., VANDERLEYDEN, J. & DE KEERSMAECKER, S. C. 2008. Genes and molecules of lactobacilli supporting probiotic action. *Microbiol Mol Biol Rev,* 72**,** 728-64.

LI, Y., CANCHAYA, C., FANG, F., RAFTIS, E., RYAN, K. A., VAN PIJKEREN, J. P., VAN SINDEREN, D. & O'TOOLE, P. W. 2007. Distribution of megaplasmids in Lactobacillus salivarius and other lactobacilli. *J Bacteriol,* 189**,** 6128-39.

MARTINO, M. E., BAYJANOV, J. R., CAFFREY, B. E., WELS, M., JONCOUR, P., HUGHES, S., GILLET, B., KLEEREBEZEM, M., VAN HIJUM, S. A. & LEULIER, F. 2016. Nomadic lifestyle of Lactobacillus plantarum revealed by comparative genomics of 54 strains isolated from different habitats. *Environ Microbiol*.

MESSAOUDI, S., MANAI, M., KERGOURLAY, G., PREVOST, H., CONNIL, N., CHOBERT, J. M. & DOUSSET, X. 2013. Lactobacillus salivarius: bacteriocin and probiotic activity. *Food Microbiol,* 36**,** 296-304.

MM, O. D., HARRIS, H. M., LYNCH, D. B., ROSS, R. P. & O'TOOLE, P. W. 2015. Lactobacillus ruminis strains cluster according to their mammalian gut source. *BMC Microbiol,* 15**,** 80.

MORARIU, V. I., SRINIVASAN, B. V., RAYKAR, V. C., DURAISWAMI, R. & DAVIS, L. S. 2009. Automatic online tuning for fast Gaussian summation. *Advances in Neural Information Processing Systems***,** 1113-1120.

MORITA, H., SHIRATORI, C., MURAKAMI, M., TAKAMI, H., KATO, Y., ENDO, A., NAKAJIMA, F., TAKAGI, M., AKITA, H., OKADA, S. & MASAOKA, T. 2007. Lactobacillus hayakitensis

sp. nov., isolated from intestines of healthy thoroughbreds. *Int J Syst Evol Microbiol,* 57**,** 2836-9.

NEVILLE, B. A. & O'TOOLE, P. W. 2010. Probiotic properties of Lactobacillus salivarius and closely related Lactobacillus species. *Future Microbiol,* 5**,** 759-74.

NILSEN, T., NES, I. F. & HOLO, H. 2003. Enterolysin A, a cell wall-degrading bacteriocin from Enterococcus faecalis LMG 2333. *Appl Environ Microbiol,* 69**,** 2975-84.

NOGUCHI, H., PARK, J. & TAKAGI, T. 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res,* 34**,** 5623-30.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res,* 27**,** 29-34.

OJALA, T., KANKAINEN, M., CASTRO, J., CERCA, N., EDELMAN, S., WESTERLUND-WIKSTROM, B., PAULIN, L., HOLM, L. & AUVINEN, P. 2014. Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics,* 15**,** 1070.

PATTEN, D. A. & LAWS, A. P. 2015. Lactobacillus-produced exopolysaccharides and their potential health benefits: a review. *Benef Microbes,* 6**,** 457-71.

PEREZ-CANO, F. J., DONG, H. & YAQOOB, P. 2010. In vitro immunomodulatory activity of Lactobacillus fermentum CECT5716 and Lactobacillus salivarius CECT5713: two probiotic strains isolated from human breast milk. *Immunobiology,* 215**,** 996-1004.

QIN, Q. L., XIE, B. B., ZHANG, X. Y., CHEN, X. L., ZHOU, B. C., ZHOU, J., OREN, A. & ZHANG, Y. Z. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol,* 196**,** 2210-5.

R CORE TEAM 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAFTIS, E. J., FORDE, B. M., CLAESSON, M. J. & O'TOOLE, P. W. 2014. Unusual genome complexity in Lactobacillus salivarius JCM1046. *BMC Genomics,* 15**,** 771.

RAFTIS, E. J., SALVETTI, E., TORRIANI, S., FELIS, G. E. & O'TOOLE, P. W. 2011. Genomic diversity of Lactobacillus salivarius. *Appl Environ Microbiol,* 77**,** 954-65.

RISSMAN, A. I., MAU, B., BIEHL, B. S., DARLING, A. E., GLASNER, J. D. & PERNA, N. T. 2009. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics,* 25**,** 2071-3.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M. A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-5.

SALVETTI, E., TORRIANI, S. & FELIS, G. E. 2012. The Genus Lactobacillus: A Taxonomic Update. *Probiotics Antimicrob Proteins,* 4**,** 217-26.

SENAN, S., PRAJAPATI, J. B. & JOSHI, C. G. 2014. Comparative genome-scale analysis of niche-based stress-responsive genes in Lactobacillus helveticus strains. *Genome,* 57**,** 185-92.

SLOVER, C. M. & DANZIGER, L. 2008. Lactobacillus: a Review. *Clinical Microbiology Newsletter,* 30**,** 23-27.

SMOKVINA, T., WELS, M., POLKA, J., CHERVAUX, C., BRISSE, S., BOEKHORST, J., VAN HYLCKAMA VLIEG, J. E. & SIEZEN, R. J. 2013. Lactobacillus paracasei comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS One,* 8**,** e68731.

SPIRIG, T., WEINER, E. M. & CLUBB, R. T. 2011. Sortase enzymes in Gram-positive bacteria. *Mol Microbiol,* 82**,** 1044-59.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

SUN, Z., HARRIS, H. M., MCCANN, A., GUO, C., ARGIMON, S., ZHANG, W., YANG, X., JEFFERY, I. B., COONEY, J. C., KAGAWA, T. F., LIU, W., SONG, Y., SALVETTI, E., WROBEL, A.,

RASINKANGAS, P., PARKHILL, J., REA, M. C., O'SULLIVAN, O., RITARI, J., DOUILLARD, F. P., PAUL ROSS, R., YANG, R., BRINER, A. E., FELIS, G. E., DE VOS, W. M., BARRANGOU, R., KLAENHAMMER, T. R., CAUFIELD, P. W., CUI, Y., ZHANG, H. & O'TOOLE, P. W. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun,* 6**,** 8322.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. 1997. A genomic perspective on protein families. *Science,* 278**,** 631-7.

TETTELIN, H., RILEY, D., CATTUTO, C. & MEDINI, D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology,* 11**,** 472-477.

WEGMANN, U., MACKENZIE, D. A., ZHENG, J., GOESMANN, A., ROOS, S., SWARBRECK, D., WALTER, J., CROSSMAN, L. C. & JUGE, N. 2015. The pan-genome of Lactobacillus reuteri strains originating from the pig gastrointestinal tract. *BMC Genomics,* 16**,** 1023.

YU, C., ZAVALJEVSKI, N., DESAI, V. & REIFMAN, J. 2011. QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Res,* 39**,** e88.

ZERBINO, D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics,* CHAPTER**,** Unit-11 5.

ZHENG, J., RUAN, L., SUN, M. & GANZLE, M. 2015a. A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology. *Appl Environ Microbiol,* 81**,** 7233-43.

ZHENG, J., ZHAO, X., LIN, X. B. & GANZLE, M. 2015b. Comparative genomics Lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. *Sci Rep,* 5**,** 18234.

Chapter IV

Application and optimization of sequence alignment protocols identifies differences in evolutionary rate across sub-clades in the genus *Lactobacillus*

In preparation as:

Application and optimization of sequence alignment protocols identifies differences in evolutionary rate across sub-clades in the genus *Lactobacillus*

Harris HMB, Claesson MJ, O'Toole PW.

TABLE OF CONTENTS

1 INTRODUCTION 149
1.1 *LACTOBACILLUS* AND EVOLUTIONARY RATE 149
1.2 SEQUENCE ALIGNMENT AND MULTIPLE SUBSTITUTION 150
1.3 AIMS OF THE STUDY 151
2 METHODS 152
3 RESULTS AND DISCUSSION 161
3.1 SEQUENCE ALIGNMENT METHODS 161
3.2 FRAME-SHIFTED ALIGNMENTS 167
3.3 RELIABLE ALIGNMENT OF IDENTICAL SEQUENCES 170
3.4 EVOLUTIONARY RATE ACROSS SUB-CLADES 171
4 CONCLUSIONS 178
5 BIBLIOGRAPHY 179

1 INTRODUCTION

1.1 *LACTOBACILLUS* AND EVOLUTIONARY RATE

Lactobacillus is a diverse, paraphyletic genus with over 247 species and subspecies according to the LPSN (List of prokaryotic names with standing in nomenclature), although numerous lactobacilli in this list have been reclassified to other genera. Lactobacilli occupy a wide range of niches including fermented meats, dairy products and the gastro-intestinal and urinary tracts of mammals (Walter, 2008). *Lactobacillus* is an interesting genus to study, not only for its functional diversity and prevalence of horizontal gene transfer (HGT), but also for the insights into biological principles such as evolutionary rate variation and selection pressure that such a diverse dataset can provide (Claesson et al., 2008).

Salvetti *et al* (in prep) generated a phylogenetic tree of 238 *Lactobacillus* species and related genera using a concatenation of 29 core ribosomal protein sequences and a maximum-likelihood approach (Figure 4.1). They used tree topology to visually identify 14 separate sub-clades ranging in size from just two species to a large sub-clade consisting of 43 species. Salvetti *et al* demonstrated robustness of these sub-clades using multiple methods of whole-genome alignment, while also conducting a functional analysis to reveal sub-clade-specific genes.

Makarova & Koonin used a molecular clock test to show that genera within the order *Lactobacillales* evolve at different rates (Makarova and Koonin, 2007). Forsdyke notes that accurate temporal calibration of evolutionary rate is difficult (Forsdyke, 2002) and other studies have suggested that absolute rather than relative rates are preferably calculated using accompanying temporal data such as fossil evidence (Jablonski and Shubin, 2015).

Relative rates of protein evolution can be estimated using information based on synonymous and non-synonymous mutations within protein-coding genes. A synonymous mutation is a single nucleotide substitution that changes the trinucleotide sequence of a codon, but that leaves the translated amino acid unchanged. A non-synonymous mutation, in contrast, will change the trinucleotide sequence and the amino acid.

Normalised values for the number of synonymous (dS) and non-synonymous (dN) substitutions capture the divergence in nucleotides that preserve and alter amino acid sequences, respectively, between two aligned, homologous, protein-coding genes. Synonymous mutations are often assumed to reflect the rate of evolution in the absence of selection (Zhang and Yang, 2015), although there is evidence that they can affect transcription and translation accuracy as well as the rate of incorrect protein folding through selective pressure acting on the 'genome phenotype' (Forsdyke, 2002). Non-synonymous mutations reflects the evolutionary rate under selective pressure where amino acid changes can alter the structural and functional properties of proteins, thereby influencing the phenotype of a genome.

The ratio of dN over dS reflects the evolutionary rate of proteins normalised for variation in mutation rate under a neutral model and can be used as a measure of the strength and type of selection pressure acting on a gene: $dN/dS = 1$ (no selection); $dN/dS > 1$ (positive selection); $dN/dS < 1$ (purifying selection) (Zhang and Yang, 2015). The dN/dS ratio is therefore a relative value of protein evolutionary rate that can be used to compare different lineages and genes.

Selective pressure can vary for each amino acid depending on its structural and functional importance (Yang, 1996). In this sense, dN/dS varies along the length of a protein and a single dN/dS value attributed to a gene reflects the average protein evolutionary rate and overall selective pressure acting on the gene.

1.2 SEQUENCE ALIGNMENT AND MULTIPLE SUBSTITUTION

Protein-coding sequences evolve as triplets of nucleotides and sequence similarity degrades more slowly at the amino acid than the nucleotide level (Abascal et al., 2010). This can lead to problems during sequence alignment, a necessary step in the calculation of values for dN and dS. Sequence alignment involves the introduction of gaps in order to preserve positional homology and, for this reason, methods that correctly align homologous codons lead to more accurate calculations of dN and dS. Alignment of genes at a nucleotide level is generally more accurate when assisted by an alignment of translated amino acid sequences acting as a template (Ranwez et al., 2011).

Correcting for multiple substitutions using algorithms such as Jukes-Cantor (Holmquist et al., 1972) can also affect the dN/dS ratio. Jukes-Cantor correction uses a simplistic model to correct for multiple substitutions, assuming that two random nucleotide sequences of the same length aligned together will share, on average, one quarter of the nucleotides from both sequences. The Jukes-Cantor formula is as follows: $3/4 \times \log (1 - 4/3 \times p)$ where p = the proportion of nucleotides not shared between two sequences. Because the value of dS is typically much higher than the value of dN (in terms of a greater number of substitutions), it will have exponentially more multiple substitutions and a higher weight for its adjusted value than that of dN, decreasing the dN/dS ratio and, potentially, changing conclusions based on selective pressure. A dN or dS value ≥ 0.75 cannot be adjusted by the JC formula because the formula assumes that sequence divergence beyond this value cannot be differentiated from an alignment of non-homologous sequences.

1.3 AIMS OF THE STUDY

This study describes the rate of protein sequence evolution of core genes using the dN/dS ratio in a large genomic dataset consisting of 227 *Lactobacillus* species and related genera. Use of the dN/dS ratio allows for interpretation of the strength and type of selection pressure acting on the core genome. The effect of sequence alignment protocol on the values of dN and dS is also described, focussing on direct nucleotide versus amino acid template alignment, pair-wise versus multiple alignment and the use of Jukes-Cantor correction for multiple substitutions. Jukes-Cantor correction was used because it makes the fewest assumptions when compared to other algorithms.

The potential for variation in protein evolutionary rate across 14 sub-clades identified by Salvetti *et al* (in prep) is investigated. The average selection pressure acting on the core genome is expected to be purifying due to selective constraints on essential genes. However, the degree to which this selection acts can vary across different lineages.

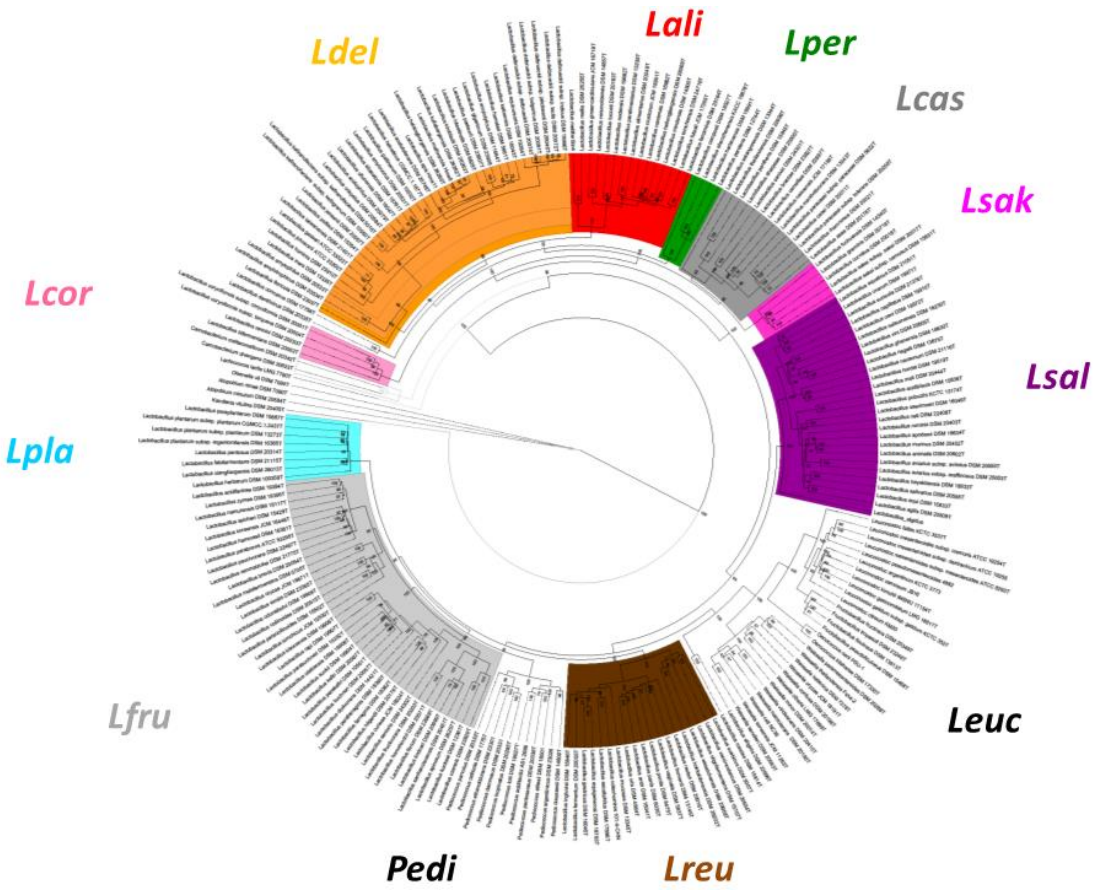


Figure 4.1: A phylogenetic tree of 238 *Lactobacillus* species and related genera (Salvetti *et al*; in prep). Ten sub-clades of *Lactobacillus* are colour-coded while *Pediococcus* and *Leuconostocaceae* are left in white. Two *Lactobacillus* sub-clades consisting of two species each have also been left in white. The out-group beneath the sub-clade Lcor is composed of *Carnobactrium*, *Atopobium*, *Kandleria* and *Lactococcus*.

2 METHODS

The Sun *et al* dataset consisting of 213 genomes (Sun et al., 2015) was supplemented with additional *Lactobacillus* genomes made available on NCBI since its publication, giving a dataset of 227 *Lactobacillus* species and related genera (comprised of 199 species and 6 genera). The 227 genomes were assigned to one of 14 separate sub-clades according to Salvetti *et al* (in prep). Table 4.1 lists all 227 genomes accompanied by four-letter abbreviations for the sub-clades to which they belong.

Genes were predicted using Glimmer3 (Delcher et al., 2007), GeneMark.HMM (Besemer et al., 2001) and MetaGene (Noguchi et al., 2006) where a gene predicted by at least one software was kept. QuartetS (Yu et al., 2011) was used to cluster gene sequences into orthologues, identifying bi-directional best hits (BHH) at thresholds of 25% identity and 30% alignment length. QuartetS output showed that the dataset of 227 genomes had a core genome of 244 genes, 166 remaining after exclusion of genes containing partial sequences (truncated 5'- and/or 3' ends).

SNAP (www.hiv.lanl.gov) was used to calculate dN, dS and dN/dS as well as JC-corrected (Jukes-Cantor) values for dN and dS to account for multiple nucleotide substitutions where dN = (number of observed non-synonymous substitutions)/(number of possible non-synonymous substitutions) and dS = (number of observed synonymous substitutions)/(number of possible synonymous substitutions). Median dN, dS and dN/dS were calculated for all pairs of homologous sequences (25,651 pair-wise alignments from 227 homologous sequences for each of 166 core genes).

Four methods were used for the alignment of 227 sequences:

1. 'nuc align': Direct nucleotide alignment of 227 sequences (separately for each of 166 core genes) using Muscle, followed by calculation of dN, dS and dN/dS for each pair of sequences using SNAP (taking a multiple alignment of 227 sequences as input).
2. 'nuc align pair': Direct pairwise nucleotide alignment of each pair of sequences (separately for each of 166 core genes) using Muscle (25,651 separate alignments), followed by calculation of dN, dS and dN/dS for each pair of sequences using SNAP (taking each pairwise alignment of two sequences as input and outputting values, one row per sequence pair).
3. 'aa align': Alignment of 227 translated amino acid sequences (separately for each of core 166 genes) using gaps as a template to replace each amino acid with its corresponding trinucleotide codon, followed by calculation of dN, dS and dN/dS for each pair of sequences using SNAP (taking a multiple alignment of 227 sequences as input).
4. 'aa align pair': Alignment of each pair of translated amino acid sequences (separately for each of 166 core genes) using gaps as a template to replace each

amino acid with its corresponding trinucleotide codon, followed by calculation of dN, dS and dN/dS for each pair of sequences using SNAP (taking each pairwise alignment of two sequences as input and outputting values, one row per sequence pair).

Lali	*Lactobacillus paralimentarius*
Lali	*Lactobacillus paralimentarius*
Lali	*Lactobacillus mindensis*
Lali	*Lactobacillus versmoldensis*
Lali	*Lactobacillus nantensis*
Lali	*Lactobacillus paralimentarius*
Lali	*Lactobacillus nodensis*
Lali	*Lactobacillus tucceti*
Lali	*Lactobacillus farciminis*
Lali	*Lactobacillus alimentarius*
Lali	*Lactobacillus kimchiensis*
Lali	*Lactobacillus mellifer*
Lali	*Lactobacillus mellis*
Lali	*Lactobacillus heilongjiangensis*
Lali	*Lactobacillus crustorum*
Lali	*Lactobacillus ginsenosidimutans*
Lali	*Lactobacillus futsaii*
Lali	*Lactobacillus crustorum*
Lcas	*Lactobacillus selangorensis*
Lcas	*Lactobacillus manihotivorans*
Lcas	*Lactobacillus selangorensis*
Lcas	*Lactobacillus pantheris*
Lcas	*Lactobacillus casei*
Lcas	*Lactobacillus rhamnosus*
Lcas	*Lactobacillus zeae*
Lcas	*Lactobacillus paracasei*
Lcas	*Lactobacillus sharpeae*
Lcas	*Lactobacillus camelliae*

Lcas	*Lactobacillus thailandensis*
Lcas	*Lactobacillus brantae*
Lcas	*Lactobacillus saniviri*
Lcas	*Lactobacillus paracasei*
Lcas	*Lactobacillus nasuensis*
Lcon	*Lactobacillus concavus*
Lcon	*Lactobacillus dextrinicus*
Lcor	*Lactobacillus coryniformis*
Lcor	*Lactobacillus bifermentans*
Lcor	*Lactobacillus coryniformis*
Lcor	*Lactobacillus rennini*
Ldel	*Lactobacillus johnsonii*
Ldel	*Lactobacillus gasseri*
Ldel	*Lactobacillus apis*
Ldel	*Lactobacillus helveticus*
Ldel	*Lactobacillus kimbladii*
Ldel	*Lactobacillus gallinarum*
Ldel	*Lactobacillus kefiranofaciens*
Ldel	*Lactobacillus amylolyticus*
Ldel	*Lactobacillus iners*
Ldel	*Lactobacillus psittaci*
Ldel	*Lactobacillus delbrueckii*
Ldel	*Lactobacillus kalixensis*
Ldel	*Lactobacillus ultunensis*
Ldel	*Lactobacillus amylovorus*
Ldel	*Lactobacillus kitasatonis*
Ldel	*Lactobacillus equicursoris*
Ldel	*Lactobacillus delbrueckii*
Ldel	*Lactobacillus delbrueckii*
Ldel	*Lactobacillus acidophilus*
Ldel	*Lactobacillus delbrueckii*
Ldel	*Lactobacillus amylovorus*
Ldel	*Lactobacillus amylophilus*
Ldel	*Lactobacillus amylotrophicus*
Ldel	*Lactobacillus jensenii*

Ldel	*Lactobacillus crispatus*
Ldel	*Lactobacillus acetotolerans*
Ldel	*Lactobacillus taiwanensis*
Ldel	*Lactobacillus floricola*
Ldel	*Lactobacillus pasteurii*
Ldel	*Lactobacillus gigeriorum*
Ldel	*Lactobacillus hominis*
Ldel	*Lactobacillus delbrueckii*
Ldel	*Lactobacillus melliventris*
Ldel	*Lactobacillus kullabergensis*
Ldel	*Lactobacillus helsingborgensis*
Ldel	*Lactobacillus kefiranofaciens*
Ldel	*Lactobacillus hamsteri*
Ldel	*Lactobacillus intestinalis*
Ldel	*Lactobacillus helveticus*
Leuc	*Leuconostoc pseudomesenteroides*
Leuc	*Leuconostoc mesenteroides*
Leuc	*Leuconostoc mesenteroides*
Leuc	*Leuconostoc mesenteroides*
Leuc	*Fructobacillus ficulneus*
Leuc	*Fructobacillus pseudoficulneus*
Leuc	*Oenococcus kitaharae*
Leuc	*Weissella minor*
Leuc	*Weissella halotolerans*
Leuc	*Weissella confusa*
Leuc	*Weissella paramesenteroides*
Leuc	*Fructobacillus fructosus*
Leuc	*Weissella viridescens*
Leuc	*Weissella kandleri*
Leuc	*Fructobacillus tropaeoli*
Leuc	*Weissella hellenica*
Leuc	*Leuconostoc kimchii*
Leuc	*Leuconostoc carnosum*
Leuc	*Weissella koreensis*
Leuc	*Weissella oryzae*

Leuc	*Leuconostoc gelidum*
Leuc	*Leuconostoc fallax*
Leuc	*Leuconostoc argentinum*
Leuc	*Leuconostoc citreum*
Leuc	*Weissella cibaria*
Leuc	*Leuconostoc gasicomitatum*
Leuc	*Weissella ceti*
Lfru	*Lactobacillus fructivorans*
Lfru	*Lactobacillus parabrevis*
Lfru	*Lactobacillus parakefiri*
Lfru	*Lactobacillus kunkeei*
Lfru	*Lactobacillus diolivorans*
Lfru	*Lactobacillus parabuchneri*
Lfru	*Lactobacillus spicheri*
Lfru	*Lactobacillus paracollinoides*
Lfru	*Lactobacillus hammesii*
Lfru	*Lactobacillus farraginis*
Lfru	*Lactobacillus parafarraginis*
Lfru	*Lactobacillus namurensis*
Lfru	*Lactobacillus acidifarinae*
Lfru	*Lactobacillus zymae*
Lfru	*Lactobacillus sunkii*
Lfru	*Lactobacillus kisonensis*
Lfru	*Lactobacillus rapi*
Lfru	*Lactobacillus otakiensis*
Lfru	*Lactobacillus odoratitofui*
Lfru	*Lactobacillus brevis*
Lfru	*Lactobacillus buchneri*
Lfru	*Lactobacillus hilgardii*
Lfru	*Lactobacillus fructivorans*
Lfru	*Lactobacillus fructivorans*
Lfru	*Lactobacillus sanfranciscensis*
Lfru	*Lactobacillus collinoides*
Lfru	*Lactobacillus fructivorans*
Lfru	*Lactobacillus kefiri*

Lfru	*Lactobacillus lindneri*
Lfru	*Lactobacillus senmaizukei*
Lfru	*Lactobacillus paucivorans*
Lfru	*Lactobacillus florum*
Lfru	*Lactobacillus similis*
Lfru	*Lactobacillus ozensis*
Lfru	*Lactobacillus senioris*
Lfru	*Lactobacillus apinorum*
Lfru	*Lactobacillus malefermentans*
Lfru	*Lactobacillus parabuchneri*
Lfru	*Lactobacillus kimchicus*
Lfru	*Lactobacillus koreensis*
Lfru	*Lactobacillus curieae*
Lfru	*Lactobacillus oryzae*
Lfru	*Lactobacillus parabrevis*
Lper	*Lactobacillus perolens*
Lper	*Lactobacillus harbinensis*
Lper	*Lactobacillus composti*
Lper	*Lactobacillus shenzhenensis*
Lpla	*Lactobacillus plantarum*
Lpla	*Lactobacillus herbarum*
Lpla	*Lactobacillus paraplantarum*
Lpla	*Lactobacillus plantarum*
Lpla	*Lactobacillus plantarum*
Lpla	*Lactobacillus pentosus*
Lpla	*Lactobacillus fabifermentans*
Lpla	*Lactobacillus xiangfangensis*
Lreu	*Lactobacillus frumenti*
Lreu	*Lactobacillus mucosae*
Lreu	*Lactobacillus ingluviei*
Lreu	*Lactobacillus oligofermentans*
Lreu	*Lactobacillus ingluviei*
Lreu	*Lactobacillus antri*
Lreu	*Lactobacillus gastricus*
Lreu	*Lactobacillus secaliphilus*

Lreu	*Lactobacillus equigenerosi*
Lreu	*Lactobacillus reuteri*
Lreu	*Lactobacillus fermentum*
Lreu	*Lactobacillus vaccinostercus*
Lreu	*Lactobacillus hokkaidonensis*
Lreu	*Lactobacillus wasatchensis*
Lreu	*Lactobacillus oris*
Lreu	*Lactobacillus suebicus*
Lreu	*Lactobacillus vaginalis*
Lreu	*Lactobacillus panis*
Lreu	*Lactobacillus pontis*
Lros	*Lactobacillus rossiae*
Lros	*Lactobacillus siliginis*
Lsak	*Lactobacillus fuchuensis*
Lsak	*Lactobacillus sakei*
Lsak	*Lactobacillus sakei*
Lsak	*Lactobacillus curvatus*
Lsak	*Lactobacillus graminis*
Lsal	*Lactobacillus mali*
Lsal	*Lactobacillus nagelii*
Lsal	*Lactobacillus acidipiscis*
Lsal	*Lactobacillus algidus*
Lsal	*Lactobacillus equi*
Lsal	*Lactobacillus acidipiscis*
Lsal	*Lactobacillus saerimneri*
Lsal	*Lactobacillus satsumensis*
Lsal	*Lactobacillus apodemi*
Lsal	*Lactobacillus ghanensis*
Lsal	*Lactobacillus hayakitensis*
Lsal	*Lactobacillus hordei*
Lsal	*Lactobacillus capillatus*
Lsal	*Lactobacillus uvarum*
Lsal	*Lactobacillus oeni*
Lsal	*Lactobacillus ruminis*
Lsal	*Lactobacillus mali*

Lsal	*Lactobacillus murinus*
Lsal	*Lactobacillus agilis*
Lsal	*Lactobacillus salivarius*
Lsal	*Lactobacillus animalis*
Lsal	*Lactobacillus vini*
Lsal	*Lactobacillus aviarius*
Lsal	*Lactobacillus aviarius*
Lsal	*Lactobacillus aquaticus*
Lsal	*Lactobacillus cacaonum*
Lsal	*Lactobacillus sucicola*
Lsal	*Lactobacillus ceti*
Lsal	*Lactobacillus pobuzihii*
Lsal	*Lactobacillus pobuzihii*
Pedi	*Pediococcus acidilactici*
Pedi	*Pediococcus claussenii*
Pedi	*Pediococcus cellicola*
Pedi	*Pediococcus stilesii*
Pedi	*Pediococcus lolii*
Pedi	*Pediococcus inopinatus*
Pedi	*Pediococcus damnosus*
Pedi	*Pediococcus parvulus*
Pedi	*Pediococcus pentosaceus*
Pedi	*Pediococcus ethanolidurans*
Pedi	*Pediococcus argentinicus*

Table 4.1: The 227 genomes of the *Lactobacillus* dataset are listed along with the sub-clades into which they are grouped (modified from Salvetti *et al*; in prep). Four-letter abbreviations are used to name each sub-clade, selecting a representative member from each in the case of *Lactobacillus* sub-clades and shortening the genus name in the case of *Leuconostocaceae* and *Pediococcus*.

3 RESULTS AND DISCUSSION

3.1 SEQUENCE ALIGNMENT METHODS

Figure 4.2 shows the percentages of the data that would be included if Jukes-Cantor correction was applied for each of the four sequence alignment methods. The number of synonymous substitutions per synonymous site has a range of between 0 and 1, all values greater than or equal to 0.75 considered as having diverged to the point of mutational saturation under the Jukes-Cantor model. Saturation here means being indistinguishable from two randomly aligned nucleotide sequences of the same length, which would disagree in three out of four bases (or 0.75). For ‘aa align’ and ‘aa align pair’ the median number of included pairs (out of 25,651) is approximately 40% while for ‘nuc align’ and ‘nuc align pair' it is much higher, showing that nucleotide alignments that use amino acid alignments as a template have a much greater number of synonymous substitutions (and a greater probability of appearing saturated at the nucleotide level).

Homologous gene datasets with high diversity can appear saturated by mutations at a nucleotide level and are difficult to distinguish from randomly aligned, non-homologous genes. These genes are still quite conserved at an amino acid level however, showing that synonymous mutations accumulate more rapidly than non-synonymous mutations, which are much more selectively constrained (Zhang and Yang, 2015). The sequence diversity across the homologous genes in *Lactobacillus* shows itself here as considerable saturation of mutations at the nucleotide level, especially for the two alignments built from amino acid templates.

The above assumptions relate to Jukes-Cantor, which assumes equal rates of transition and transversion as well as equal proportion of the four nucleotides (Holmquist et al., 1972). Other models such as Kimura (Kimura, 1980) may not lead to saturation at the same level of nucleotide divergence due to the relaxed assumption that transitions can have different rates to transversions. Jukes-Cantor may not be a suitable method for correcting for multiple substitutions at this level of sequence diversity.

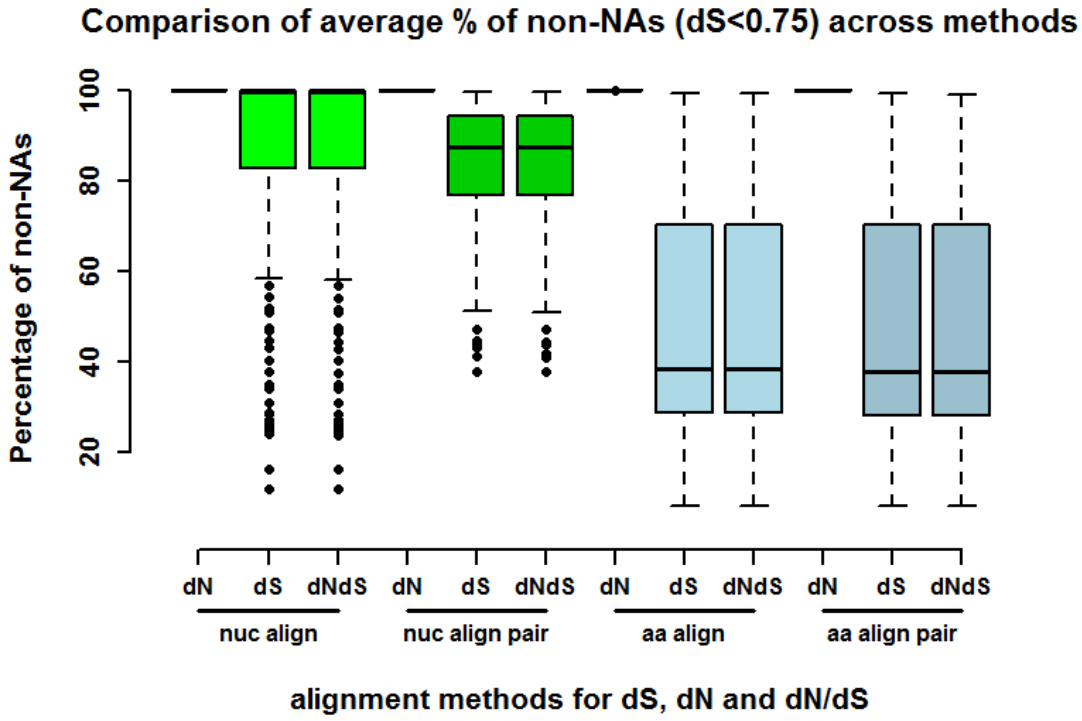


Figure 4.2: The number of NAs generated after Jukes-Cantor correction differs considerably across four sequence alignment methods. An uncorrected value becomes NA ('not applicable' as output from the software) when dS >= 0.75. The median values of 166 core genes are shown and boxplots are generated from 25,651 pair-wise comparisons involving 227 genomes. The percentage of NAs for dN/dS is equal to that of dS for each method because dN/dS cannot be computed without a numerical value for dS. Labels for the four alignment methods are assigned and explained in Methods.

Figures 4.3 and 4.4 show that aligning nucleotide sequences directly leads to fewer predicted synonymous mutations but more non-synonymous mutations. When aligning nucleotide sequences directly, Muscle increases percent identity and alignment score at the cost of introducing unnecessary gaps that misalign homologous codons. This is less an issue of multiple alignment and more to do with ignoring biological information on how sequences evolve, in this case the selective pressure acting on codons. The literature suggests that values of dN and dS (and hence dN/dS) are more accurate for 'aa align' and 'aa align pair', both of which use amino acid alignments as templates for constructing correctly aligned codons at the nucleotide level (Abascal et al., 2010).

Interestingly, local sequence aligners like BLAST would never detect distant homology at the nucleotide level because they need to match similar k-mers between two sequences (Altschul et al., 1990) and two distantly related homologous genes might have no similar sequence regions at the nucleotide level.

Even though all four methods give the same conclusion of a core genome under purifying selection pressure (Figure 4.5), 'nuc align' and 'nuc align pair' have significantly different dN/dS from each other and from 'aa align' and 'aa align pair', which lead to an interpretation of greater selective constraint acting on core genes when amino acid-based alignment methods are used.

The reason for the difference between 'nuc align' and 'nuc align pair' is likely due to methodological differences in introducing gaps in a multiple alignment of 227 sequences compared to aligning sequences two at a time. Pair-wise sequence alignment is generally more accurate and the lower dN/dS value of 'nuc align pair' (closer to 'aa align' and 'aa align pair') supports this.

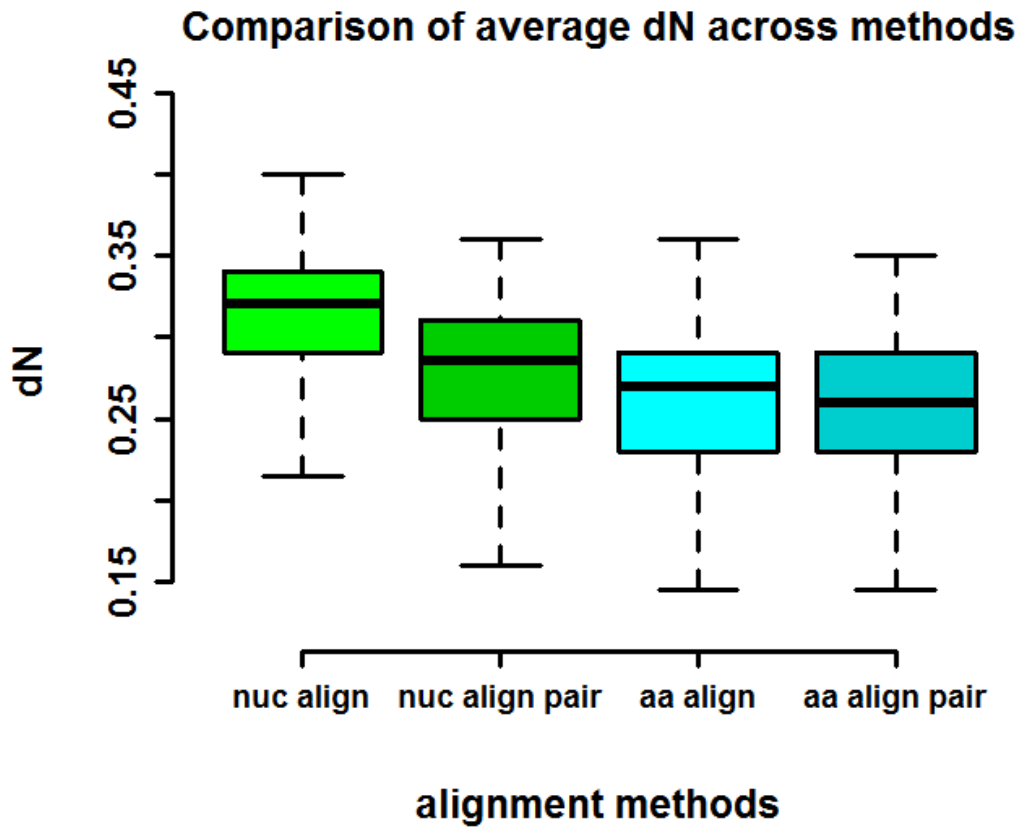


Figure 4.3: Average dN values are significantly higher when nucleotide sequences are aligned directly. The median values of 166 core genes are shown and boxplots are generated from 25,651 pair-wise comparisons involving 227 genomes. Labels for the four sequence alignment methods are assigned and explained in Methods. Outliers are excluded, but all four methods have a minimum value of 0 and maximum values of 0.4, 0.36, 0.36 and 0.35 from left to right.

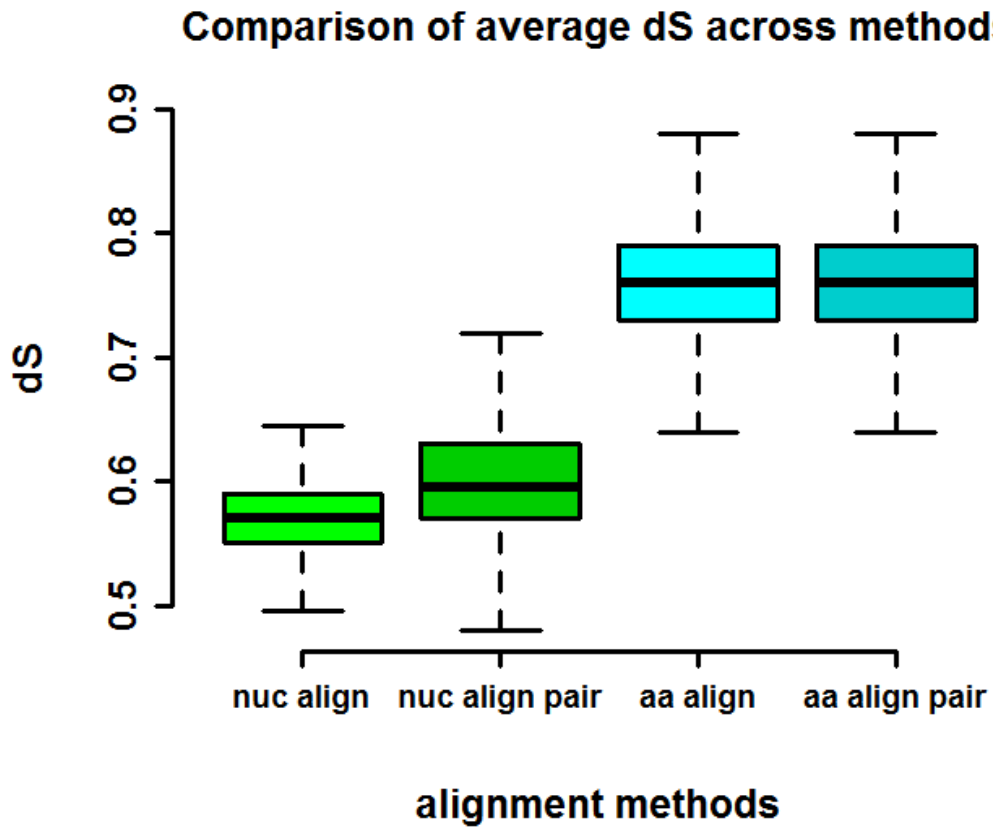


Figure 4.4: Average dS values are significantly lower when nucleotide sequences are aligned directly. The median values of 166 core genes are shown and boxplots are generated from 25,651 pair-wise comparisons involving 227 genomes. Labels for the four sequence alignment methods are assigned and explained in Methods. Outliers are excluded, but all four methods have a minimum value of 0 and maximum values of 0.72, 0.82, 0.99 and 0.99 from left to right.

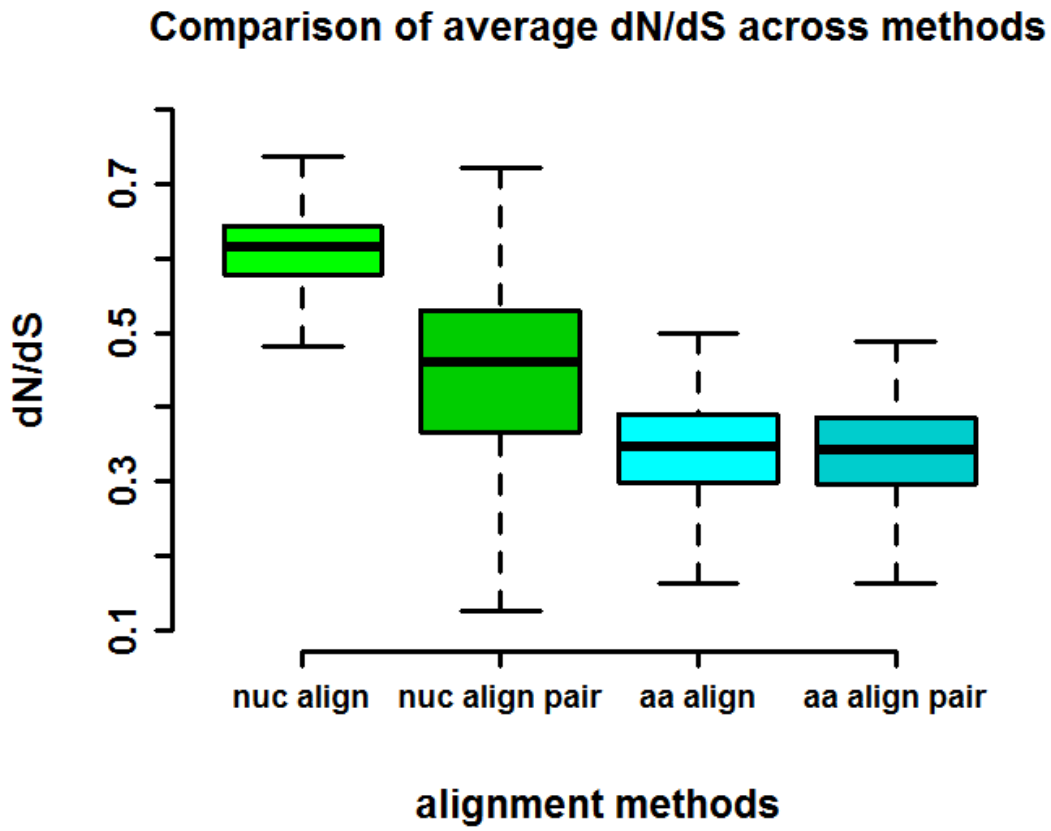


Figure 4.5: Average dN/dS ratios are significantly higher when nucleotide sequences are aligned directly. The median values of 166 core genes are shown and boxplots are generated from 25,651 pair-wise comparisons involving 227 genomes. Labels for the four sequence alignment methods are assigned and explained in Methods. Outliers are excluded, but all four methods have a minimum value of 0 and maximum values of 0.74, 0.72, 0.5 and 0.49 from left to right.

3.2 FRAME-SHIFTED ALIGNMENTS

The ‘nuc align’ alignment method took 227 sequences as input and output a multiple alignment for each of 166 core genes. For this reason, if the alignment for a particular set of homologous sequences (representing a core gene) was shifted out-of-frame, the calculation of dN and dS would be out-of-frame for all 227 sequences. In contrast, the ‘nuc align pair’ method aligned sequences one pair at a time so in-frame and out-of-frame alignments could be counted as each alignment occurred.

Figures 4.6 and 4.7 show the percentage of out-of-frame pairs (out of 25,651) for each of 166 core genes (for ‘nuc align pair’) and the average dN/dS ratio for core genes in-frame and out-of-frame (for ‘nuc align’), respectively. The number of out-of-frame alignments varies considerably across core genes ranging from close to 0% to almost 50%, suggesting that the degree of sequence conservation also varies considerably because many more gaps inserted into an alignment reflect sequence divergence and lead to a higher probability of incorrect positional homology (Abascal et al., 2010).

Frame-shifts lead to higher dN values because they mis-align homologous codons, increasing the number of non-synonymous substitutions. Frame-shifts lead to lower dS because greater sequence similarity at the nucleotide level is achieved at the expense of introducing additional gaps. The overall effect is a higher dN/dS ratio and an interpretation of higher positive selective pressure acting on core genes. Frame-shifts are not a problem for amino acid alignments because codons are treated as a single unit (i.e. the amino acid), although it is still possible that non-homologous codons can be aligned, leading to less accurate calculations of dN and dS.

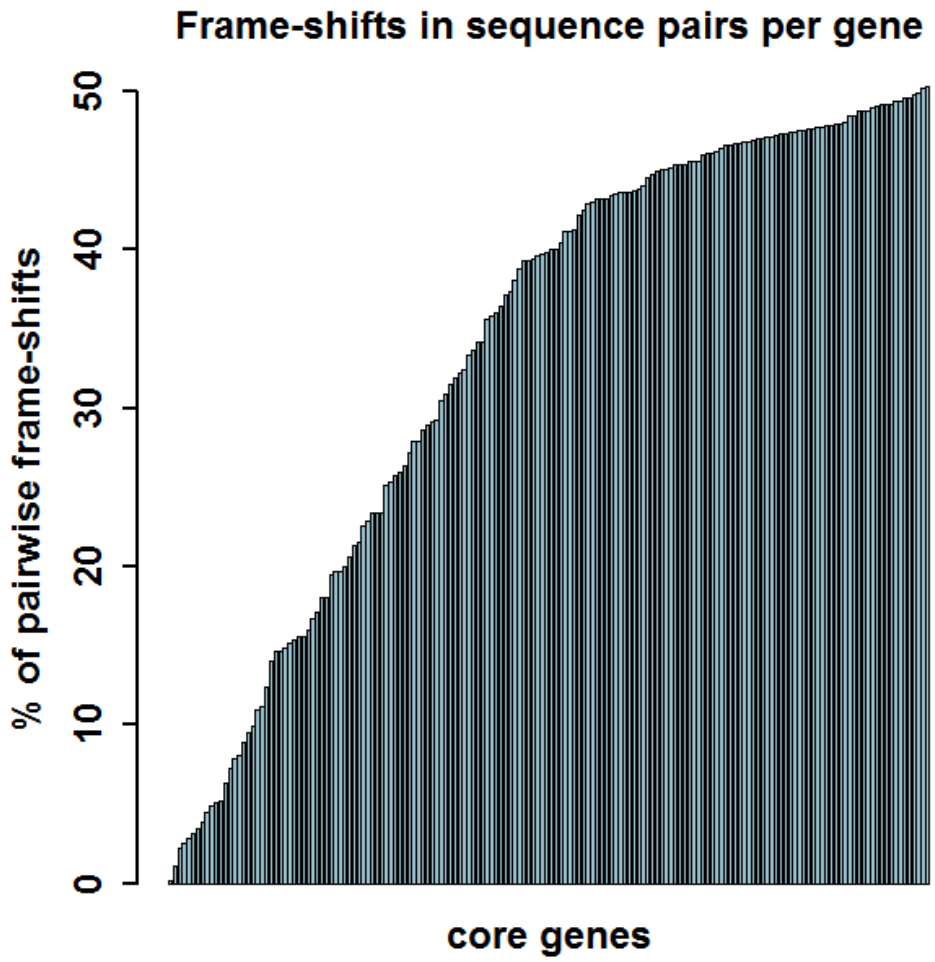


Figure 4.6: Percentage of pair-wise nucleotide alignments per gene that were frame-shifted due to addition of one or more incorrect gaps ranges from less than 1% to over 50%. Pair-wise nucleotide sequence alignments that did not have a length equal to a multiple of three (unlike the unaligned input sequences) were counted for each of 166 core genes and expressed as a percentage of 25,651 pair-wise alignments (for 'nuc align pair'). Genes are ordered according to increasing percentage of frame-shifts.

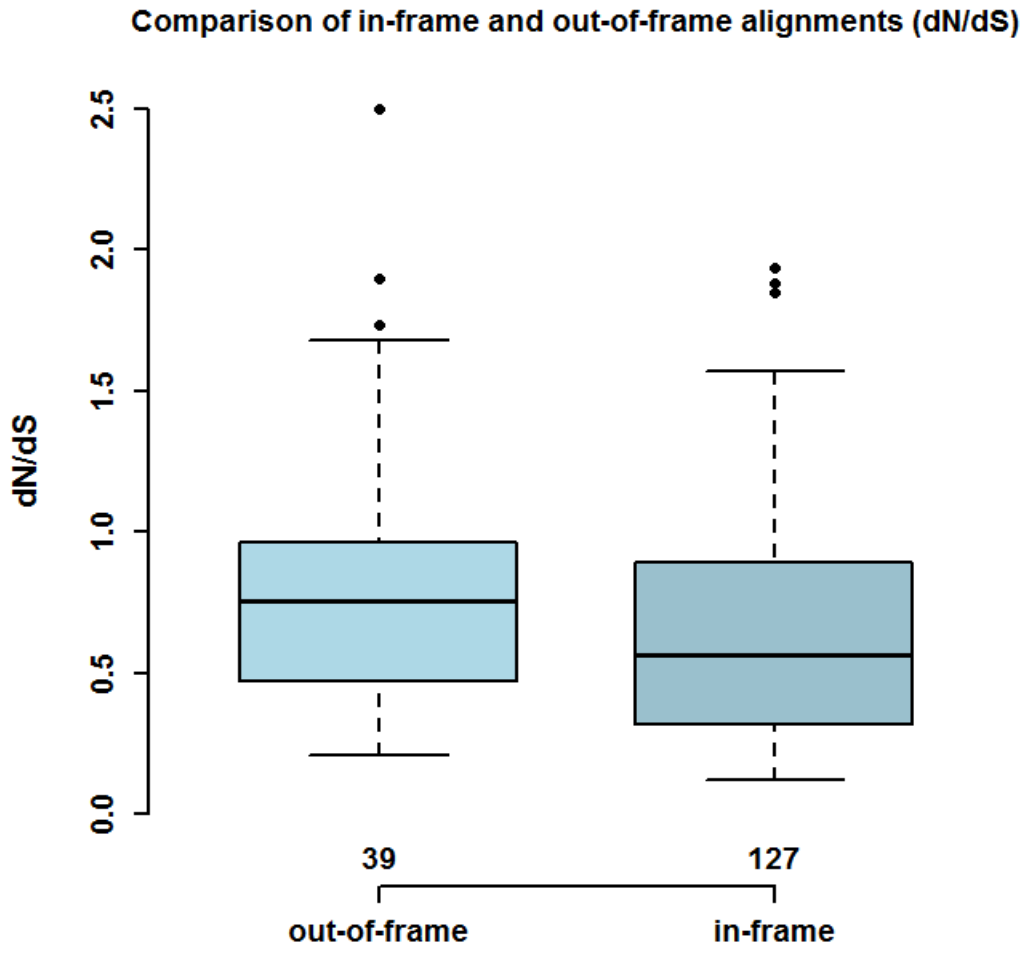


Figure 4.7: Average dN/dS ratios are significantly higher for genes where the multiple alignments of 227 sequences lead to a frame-shift somewhere along its length (for 'nuc align'). Boxplots are generated from 166 core genes (39 out-of-frame and 127 in-frame) where dN/dS for each gene is the median average of 25,651 pair-wise comparisons of 227 sequences.

3.3 RELIABLE ALIGNMENT OF IDENTICAL SEQUENCES

For each core gene in this dataset, a minority were identical. Table 4.2 shows a comparison of all-versus-all BLAST for each core gene with SNAP results of dN and dS. The relevant data are only those sequences with 100% alignment over their full length (for BLAST) and dN and dS equal to zero (for SNAP), both results indicating a protein evolutionary rate of zero.

BLAST is a local aligner, but for identical sequences and with masking of repetitive regions turned off, it will align the sequences over their full length and identify all pairs where dN and dS should equal zero. Assuming that multiple alignments are always accurate for identical sequences, calculation of dN and dS should always give values of zero. Table 4.2 shows that this is not the case, pair-wise alignments always agreeing with BLAST while multiple alignments sometimes failing to align identical sequences correctly, which can be seen from correlation values of less than one. These results suggest that multiple sequence alignments of divergent homologous sequences can lead to the mistaken identification of mutations in the minority of identical sequences present when a large number of sequences are aligned (227 in this case).

Aligning two sequences at a time from a total of 227 in a pair-wise manner, followed by calculation of dN and dS, identifies all identical pairs and agrees with BLAST results. This is true for both ‘nuc align pair’ and ‘aa align pair’. These results encourage performing sequence alignment in a pair-wise manner for analyses involving evolutionary rate and selection pressure. The main advantage of multiple alignment over pair-wise alignment in these cases is time, a factor that should be balanced with the accurate generation and interpretation of results.

	zero count	nuc align	aa align	nuc align pair	aa align pair
zero count	1	0.884768	0.958548	1	1
nuc align	0.884767927	1	0.903647	0.884767927	0.884767927
aa align	0.958548076	0.903647	1	0.958548076	0.958548076
nuc align pair	1	0.884768	0.958548	1	1
aa align pair	1	0.884768	0.958548	1	1

Table 4.2: A Spearman correlation of the number of pair-wise comparisons consisting of identical sequences (out of 25,651) for each gene shows that multiple alignments can incorrectly align identical sequences in a minority of cases. BLAST was used to align all 227 sequences for each of the 166 core genes against each other. The number of identical pair-wise alignments is compared with the number of pair-wise comparisons where both dN and dS equal 0 (representing zero mutations between identical sequences) for each of the four sequence alignment methods. The number of identical sequences per gene ranges from 6 to 38 with a median of 12.

3.4 EVOLUTIONARY RATE ACROSS SUB-CLADES

In Figures 4.8 and 4.9, it can be seen that ‘nuc align’ shows similar trends in dN/dS across sub-clades whether Jukes-Cantor correction was applied or not, with Leuconostocaceae having reduced purifying selection (increased dN/dS) acting on the core genome in both cases. *Leuconostoc* has been shown to evolve faster than *Pediococcus* in a previous study and at least one strain of *Oenococcus oeni* apparently lacks MutL and MutS (Makarova and Koonin, 2007), a result supported by the lack of strong BLAST hits to mutS and mutL in *O. oeni* ATCC-BAA 1163 in this dataset.

Jukes-Cantor correction decreases average dN/dS across the sub-clades by increasing the value of dS more than dN due to its non-linear correction for multiple substitutions. This means that a higher value will be corrected to a proportionally higher value by Jukes-Cantor.

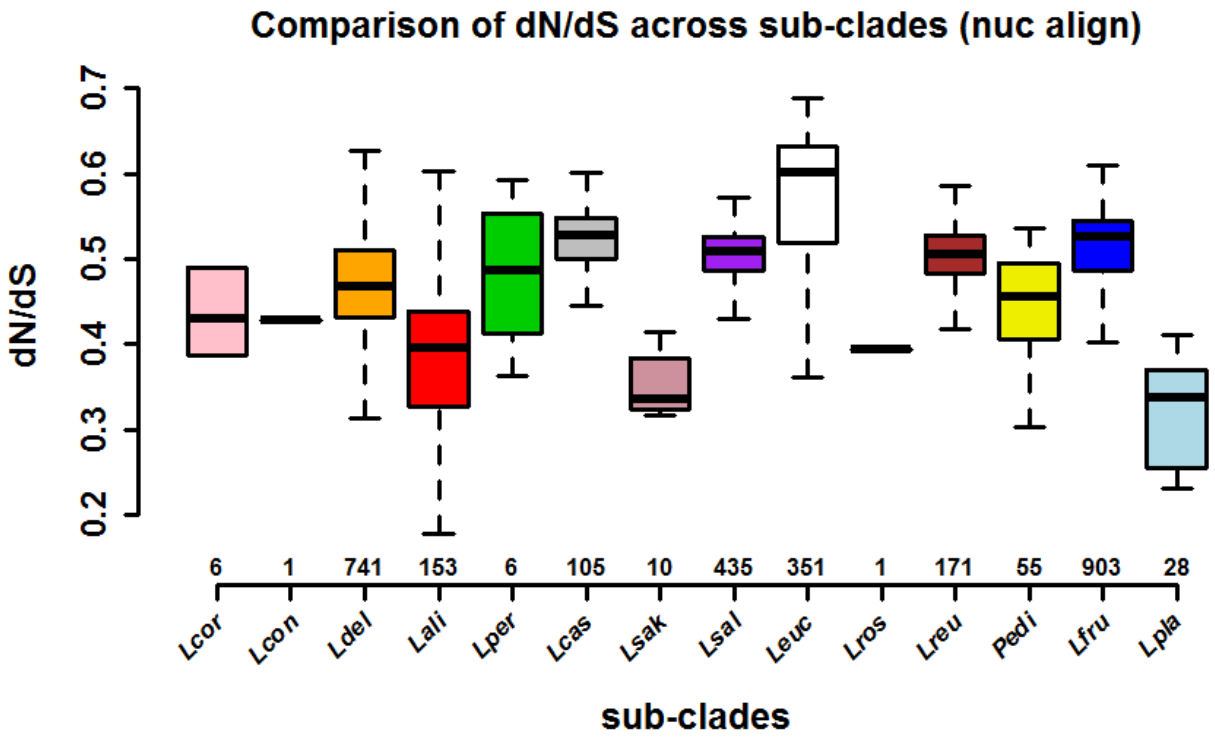


Figure 4.8: The dN/dS ratio varies significantly across 14 sub-clades for 'nuc align'. The dataset of 227 genomes was divided into 14 sub-clades as described in Methods. The median values of 166 core genes are shown and boxplots are generated from the number of pair-wise comparisons displayed over the x-axis labels. Labels for sub-clades are four-letter abbreviations of a representative member of each group, the full membership of which is listed in Methods (Table 4.1). The order of the boxplots follows the clock-wise order of sub-clades in the phylogenetic tree described and displayed in Methods (Figure 4.1). Outliers are excluded from this figure, but do not change the scale of the y-axis. Jukes-Cantor correction was not applied to the values in this figure.

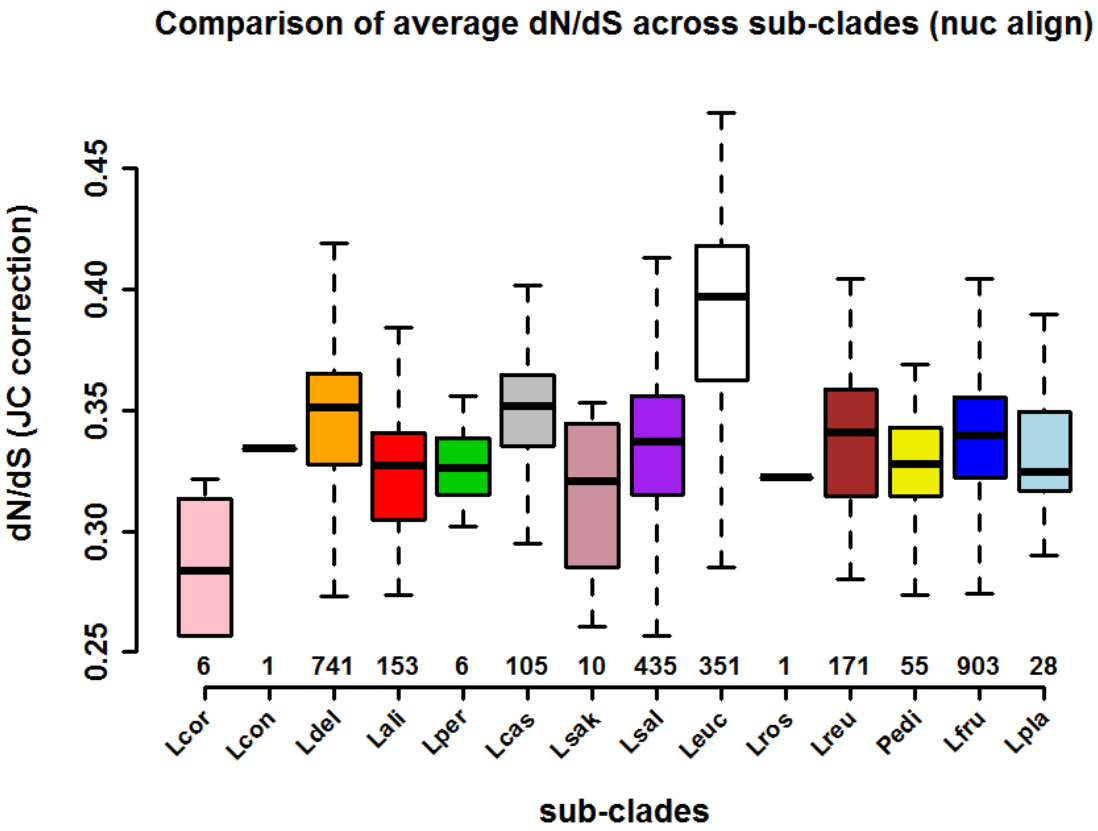


Figure 4.9: Trends across sub-clades for the JC-corrected dN/dS ratio largely agree with the uncorrected values shown in figure 4.8, but average values are consistently lower. Methods are identical to those for Figure 4.8. Jukes-Cantor correction was applied to the values in this figure.

Figure 4.10 shows that despite lower average dN/dS ratios across sub-clades, the relative values remain consistent between sub-clades for 'nuc align' and 'aa align'. These values are median averages of a very large number of pair-wise comparisons. It is likely that trends would be in less agreement for a single gene in a smaller dataset across alignment methods.

There is a tendency for sub-clades with more genomes (and therefore more pair-wise comparisons) to have more relaxed purifying selection pressure acting on the core genome, suggesting that group size biases the dN/dS ratio towards higher values. This is not necessarily true however, as dN/dS is already a normalised value

that reflects the evolutionary rate of protein sequences under selection, accounting for possible variations in the neutral mutation rate across sub-clades (using dS) (Zhang and Yang, 2015). An alternative explanation is that sub-clades may be under different degrees of positive selection pressure from their respective niches, but without additional temporal information for the common ancestors of each sub-clade, interpretation of the relationship between sub-clade size and dN/dS ratio is difficult.

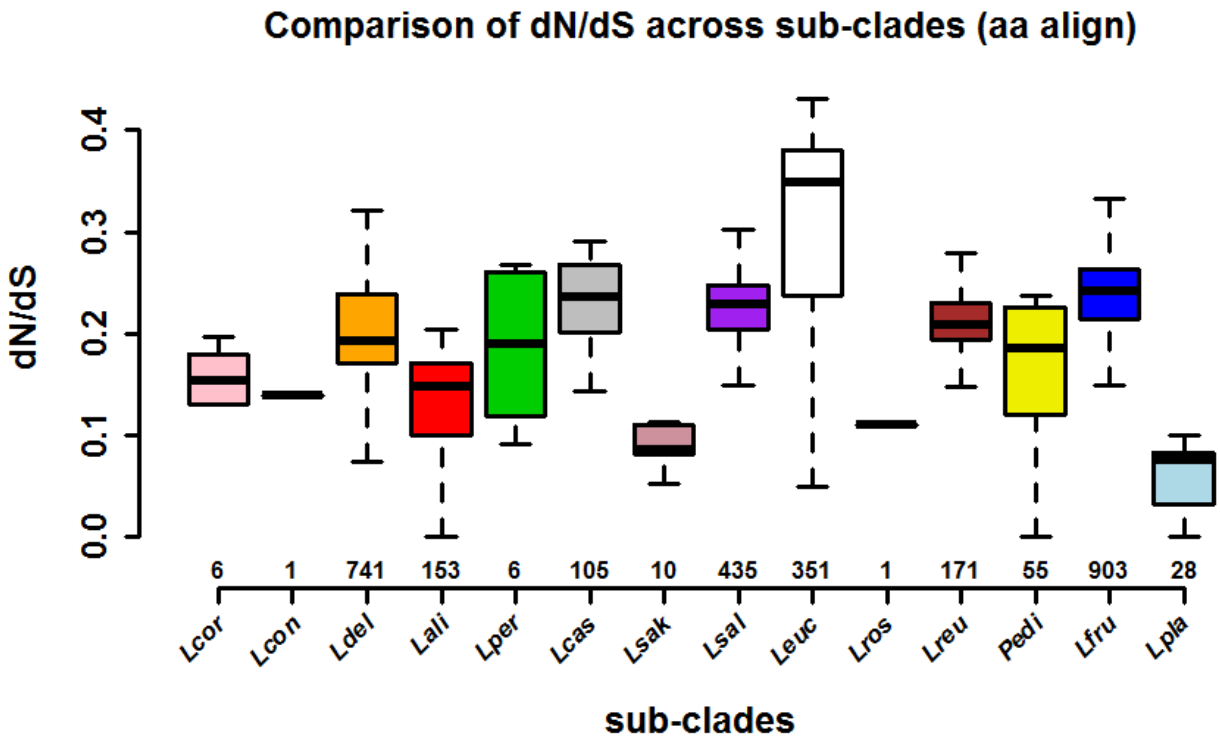


Figure 4.10: The dN/dS ratio across sub-clades for amino-acid based alignment resembles that for direct nucleotide-based alignment. Methods are identical to those for Figure 4.8. Jukes-Cantor correction was not applied to the values in this figure.

Amino acid-based alignment (‘aa align’) and direct nucleotide alignment (‘nuc align’) disagree on the role of positive selection pressure acting on core genes with Figure 4.12 showing numerous genes across sub-clades with dN/dS >1. The alignment methods, ‘nuc align’ and ‘nuc align pair’, give higher dN, lower dS and therefore a higher average dN/dS ratio when compared to ‘aa align’ and ‘aa align pair’. It is not surprising therefore that average dN/dS is higher in Figure 4.12 than in Figure 4.11 across the sub-clades, but the difference is even more exaggerated. In previous figures, dN/dS was averaged over genes while gene values are displayed individually in these two figures, showing the variation that exists in purifying selection from gene to gene.

Figure 4.11 is likely to be more correct both because core genes, on average, are very probably under purifying selection (Bohlin et al., 2017) and because amino acid-based alignments take the evolution of sequences as nucleotide triplets into account, which leads to a more reliable alignment of homologous codons (Ranwez et al., 2011)

The results suggest that, as more specific analyses are undertaken involving dN/dS, the choice of alignment method becomes more important in arriving at the correct conclusions. In this case, the average predicted protein evolutionary rate of each sub-clade becomes separated out across 166 core genes (when comparing Figures 4.8 and 4.10 with Figures 4.11 and 4.12), leading to a greater range of values for dN/dS.

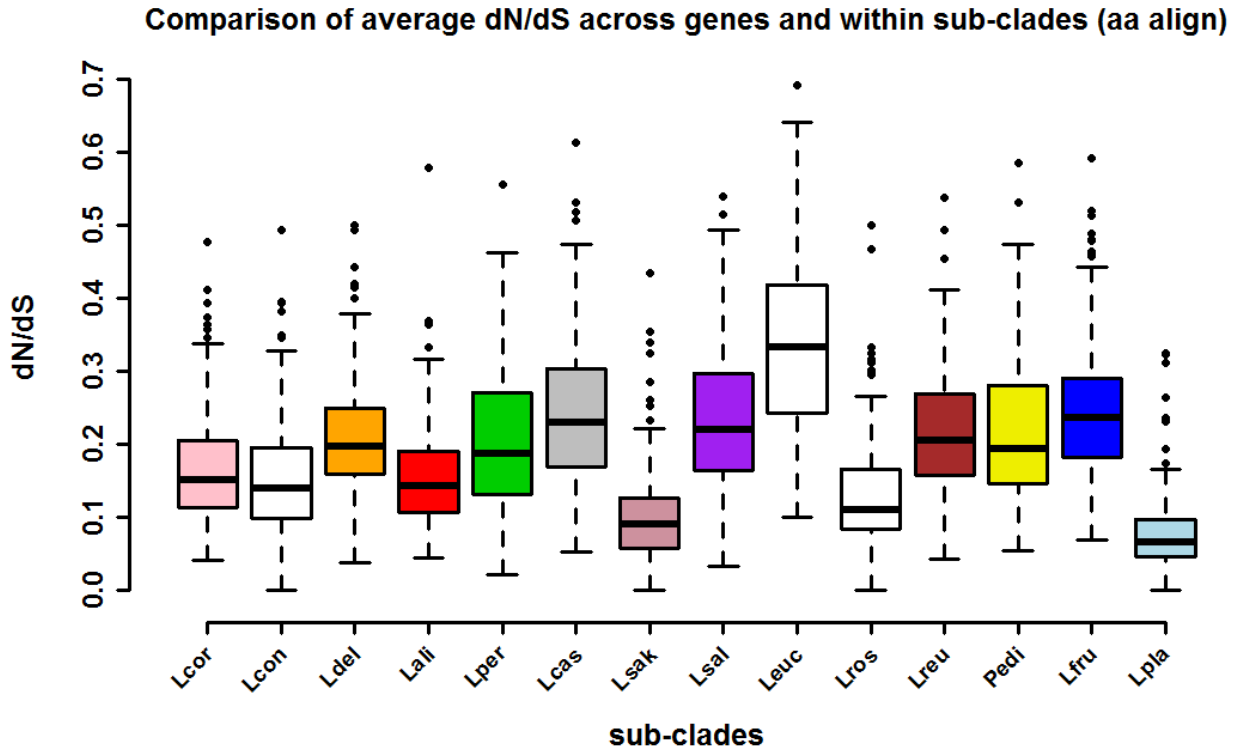


Figure 4.11: Average dN/dS for 166 core genes varies significantly across 14 sub-clades for amino acid-based alignment and all genes are dominated by purifying selection pressure. The median value of the number of pair-wise comparisons displayed on the x-axis in Figure 4.8 is shown for each gene. Boxplots for each sub-clade are constructed from 166 core-gene values. Jukes-Cantor correction was not applied to the values in this figure.

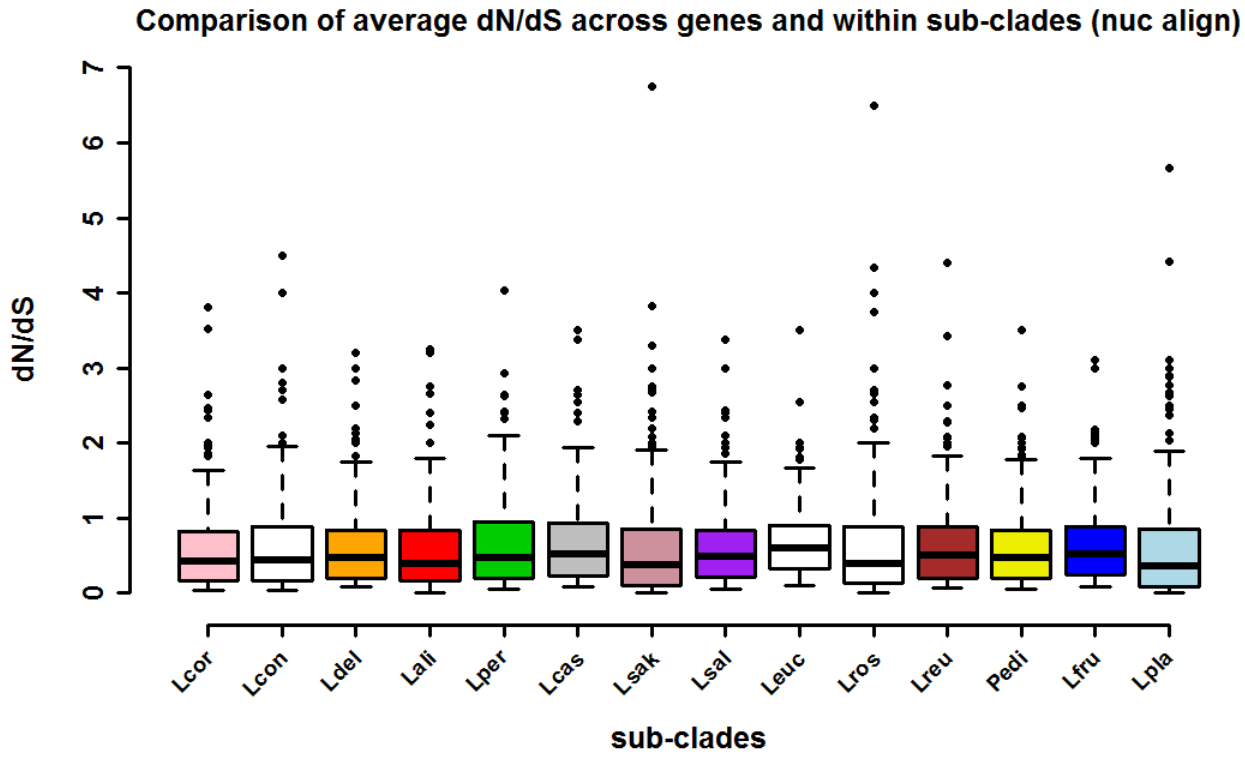


Figure 4.12: Average dN/dS for 166 core genes varies significantly across 14 sub-clades for nucleotide-based alignment and the dominant selective pressure acting on some genes appears to be positive. The median value of the number of pair-wise comparisons displayed on the x-axis in Figure 4.8 is shown for each gene. Boxplots for each sub-clade are constructed from 166 core-gene values. Jukes-Cantor correction was not applied to the values in this figure.

4 CONCLUSIONS

For a large, phylogenetically diverse dataset, a considerable proportion of homologous sequence comparisons can show saturation at the nucleotide level. This conclusion is, however, based on the Jukes-Cantor model and other models that make more realistic assumptions and account for nucleotide codon triplets may lead to different conclusions.

Sub-clades across the paraphyletic *Lactobacillus* genus vary in the protein evolutionary rate of their core genes, possibly due to differences in selection pressure by the environment or differences in the efficiency of DNA repair mechanisms.

Different sequence alignment methods give significantly different values of dN, dS and dN/dS, choice of method being important when interpreting how these values reflect evolutionary rate and strength of selection pressure.

The zoomed-out approach of this study to comparing average dN, dS and dN/dS across a large dataset very probably made results more robust to the effects of different sequence alignment methods and Jukes-Cantor correction. It would be interesting to observe the results of similar studies on a subset of the data, perhaps focussing on one or several genes within a sub-clade. The average selection pressure of the core genome is purifying, but results of these analyses on two subsets of genes, one under purifying and one under positive selection pressure, might reveal interesting differences in how software tools and algorithms behave in these two scenarios.

Also, the comparison of multiple sequence aligners as well as several algorithms for multiple substitution correction would be a relevant extension to this chapter because varying assumptions can and do lead to different evolutionary conclusions.

5 BIBLIOGRAPHY

ABASCAL, F., ZARDOYA, R. & TELFORD, M. J. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. ***Nucleic Acids Res,*** **38, W7-13.**

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. ***J Mol Biol,*** **215, 403-10.**

BESEMER, J., LOMSADZE, A. & BORODOVSKY, M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. ***Nucleic Acids Res,*** **29, 2607-18.**

BOHLIN, J., ELDHOLM, V., PETTERSSON, J. H., BRYNILDSRUD, O. & SNIPEN, L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. ***BMC Genomics,*** **18, 151.**

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE, P. W. 2008. Lactobacillus phylogenomics--towards a reclassification of the genus. ***Int J Syst Evol Microbiol,*** **58, 2945-54.**

DELCHER, A. L., BRATKE, K. A., POWERS, E. C. & SALZBERG, S. L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. ***Bioinformatics,*** **23, 673-9.**

FORSDYKE, D. R. 2002. Selective pressures that decrease synonymous mutations in Plasmodium falciparum. ***Trends Parasitol,*** **18, 411-7.**

HOLMQUIST, R., CANTOR, C. & JUKES, T. 1972. Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids. ***J Mol Biol,*** **64, 145-61.**

JABLONSKI, D. & SHUBIN, N. H. 2015. The future of the fossil record: Paleontology in the 21st century. ***Proc Natl Acad Sci U S A,*** **112, 4852-8.**

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. ***J Mol Evol,*** **16, 111-20.**

MAKAROVA, K. S. & KOONIN, E. V. 2007. Evolutionary genomics of lactic acid bacteria. ***J Bacteriol,*** **189, 1199-208.**

NOGUCHI, H., PARK, J. & TAKAGI, T. 2006. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. ***Nucleic Acids Res,*** **34, 5623-30.**

RANWEZ, V., HARISPE, S., DELSUC, F. & DOUZERY, E. J. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. ***PLoS One,*** **6, e22594.**

SUN, Z., HARRIS, H. M., MCCANN, A., GUO, C., ARGIMON, S., ZHANG, W., YANG, X., JEFFERY, I. B., COONEY, J. C., KAGAWA, T. F., LIU, W., SONG, Y., SALVETTI, E., WROBEL, A., RASINKANGAS, P., PARKHILL, J., REA, M. C., O'SULLIVAN, O., RITARI, J., DOUILLARD, F. P., PAUL ROSS, R., YANG, R., BRINER, A. E., FELIS, G. E., DE VOS, W. M., BARRANGOU, R., KLAENHAMMER, T. R., CAUFIELD, P. W., CUI, Y., ZHANG, H. & O'TOOLE, P. W. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. ***Nat Commun,*** **6, 8322.**

WALTER, J. 2008. Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. ***Appl Environ Microbiol,*** **74, 4985-96.**

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. ***Trends Ecol Evol,*** **11, 367-72.**

YU, C., ZAVALJEVSKI, N., DESAI, V. & REIFMAN, J. 2011. QuartetS: a fast and accurate algorithm for large-scale orthology detection. ***Nucleic Acids Res,*** **39, e88.**

ZHANG, J. & YANG, J. R. 2015. Determinants of the rate of protein sequence evolution. ***Nat Rev Genet,*** **16, 409-20.**

Chapter V

General discussion and future perspectives

Charles Darwin devoted a chapter of his book *On the Origin of Species* to the artificial selection of species by man. It described the variation present in domesticated plants and animals, detailing the ways that humans deliberately or accidentally selected for particular traits, usually traits that benefitted the humans involved. Darwin chose to explain selection in nature using the analogy of conscious selection by man, showing that the environment also leads to differential survival and reproduction of phenotypes. He was well aware of the power that earlier civilisations had, not only in reshaping the abiotic conditions that surrounded them, but also in moulding local species to suit human needs. From the development of animal husbandry and plant breeding that led to the expansion of the first settled populations to the diversification of the rock dove at the whims of generations of pigeon fanciers, Darwin described how people's early, basic knowledge of heredity gave them the ability to exploit the variation inherent in biological resources, selecting the most favourable varieties and increasing their economic value over time (Darwin, 1859).

Early attempts to control the reproduction of species in order to gain from the resources they provide can be viewed as the origin of biotechnology. As crude and non-scientific as they were, these endeavours reflect the human capacity to recognise patterns in the environment and to redirect those patterns for the benefit of human survival and reproduction. It is the ability to manipulate our environment, to make it more amenable to our needs, rather than to simply struggle to adapt to its changing conditions, that has led to advances in knowledge about the species around us, from the first civilisations to the present day.

The use of macroscopic, multicellular organisms such as species of livestock and cereal to support dense communities of people is an impressive display of human ingenuity, laying the foundation for the development of complex societies through division of labour and specialisation (Violatti, 2014). It is in the microscopic world, however, that we are likely to see the greatest biotechnological innovations and insights.

Microbes represent a large portion of the genetic diversity of life, playing essential biogeochemical roles on a global scale, including those of carbon, nitrogen and phosphorous cycling (Wang et al., 2017). They colonise our bodies in a mutualistic relationship that, in the case of the gastro-intestinal tract, allows humans to absorb otherwise indigestible carbohydrates and additional bacterial products of

metabolism such as short-chain fatty acids (Flint et al., 2012). Microbes also represent one of our biggest challenges, causing an alarming array of diseases that are evolutionarily adopting their former infectivity because of the rise of strains of bacterial pathogens resistant to antibiotics, which is currently one of the biggest threats to global health and food security (WHO, 2017).

Humans have a long history of exploiting microbes too. From the first Neolithic farmers getting drunk on fermented beverages (Charles and Durham, 1952) to the health-conscious people of modern society who drink probiotics as part of their daily diet, we have benefited from microscopic organisms without understanding how they function or even knowing, in earlier times, that they exist at all. It was with the discovery of bacteria by Antonie Van Leeuwenhoek (Lane, 2015), followed by the pioneering work of people such as Robert Koch and Louis Pasteur (Hook, 2011), that the study of microorganisms as a scientific discipline was established, turning the microbial benefits and dangers that we experience away from the beliefs and superstitions of the day and grounding them in quantitative, empirical observations and experiments.

In 2017, Microbiology faces a whole range of challenges that could hardly be predicted in Pasteur and Koch's time and that did not exist before the advent of DNA sequencing. Bioinformatics and sequencing technology give us the tools to answer questions that were previously beyond reach. How diverse is the complex microbial ecology of the gut? What is its functional capacity? How do the species interact under different environmental conditions such as changes in diet and age, affecting human health and wellbeing in the process (Claesson et al., 2012)? Similar questions could be asked for microbes inhabiting the soil, the oceans, our food and even the atmosphere as described by a recent paper on the microbial communities of clouds (Amato et al., 2017).

The focus is no longer on the genetic and phenotypic properties of one or a few strains in isolation, but on entire microbial communities. This does not reflect changing interests so much as changing technological and computational capabilities. In similar fashion, the modern perspective of a bacterial species is shifting from the defining properties of the type strain to the functional variation encoded in the pan-genome, a transition that is especially important for microbes with strain-dependent pathogenicity like *E. coli* (Rasko et al., 2008) or that have

strain-specific commercial properties like many Lactic Acid Bacteria (Campana et al., 2017). The emphasis is very much on diversity.

The ability to generate huge amounts of data that capture the diversity of microbial communities is an impressive development in modern biological research. It brings with it a responsibility for current and future generations of biologists to efficiently manage, store and analyse these data. A recent study led by Professor Rob Knight in collaboration with the Earth Microbiome Project presented a meta-analysis of hundreds of microbial community samples from around the world, aiming to give a more complete characterization of microbial life on Earth (Thompson et al., 2017). They highlight a growing awareness of the importance and diversity of microbes, stressing that this is in stark contrast with our limited understanding, an obstacle that is partly due to a lack of standardised protocols and analytical frameworks, but also related to the sheer phylogenetic and functional diversity of species.

Similar challenges exist in Comparative Microbial Genomics where the number of available sequenced genomes is growing exponentially and the true extent of genomic diversity within species is only beginning to be charted. The phylogenomic complexity of taxa is increasingly viewed as an exciting area of research as the role of horizontal gene transfer in microbes is better understood. Back in 2003, a team at Lawrence Livermore National Laboratory was given the task of improving the computational tools necessary for fast and efficient DNA diagnostics for pathogen detection (Chain et al., 2003). The conclusion of their assessment of software at the time was that the selection of appropriate tools can have a large effect on both the quality of results and on the effort required to reach those results. They list the accuracy of results on gene function, gene regulation, gene networks, phylogenetic studies and other aspects of evolution as depending on accurate analytical methods. Fourteen years later, another review on bioinformatic platforms for comparative genomics echoes similar conclusions, emphasising the demand for fast and automated approaches to keep pace with the rapid increase in available microbial genomes (Yu et al., 2017).

The necessity for computational speed and automation is obvious enough given the increasing size of biological datasets. Appropriate software tools and better reference databases are also essential if the biology represented by sequence data is to be fully understood. The result is more powerful computers, more sophisticated algorithms, more experienced bioinformaticians and more efficient storage of digital

data. The future of the scientific enterprise will not depend on these things alone; there are other factors, more subtle perhaps, but every bit as important that must be accounted for if all these sequencing data and their supporting analytical tools and tool-makers are to achieve their potential. It goes back to the heart of the Scientific Revolution, grounded in collaboration and the importance of asking the right questions.

The origin of a study begins with a hypothesis, a question or series of questions sometimes clearly defined, sometimes more exploratory. It often requires a carefully crafted experimental design in order to bring that question closer to a possible answer; in the field of bioinformatics, this starts with sequencing the right DNA. This suggestion holds considerable merit. An experiment involving sequencing and its accompanying analysis takes time, expertise and money. The results are generated, interpreted and published where they can be read and reviewed and criticised. It does not stop there, however. By making the sequence data publicly available, other researchers, perhaps even those without the funding for sequencing projects of their own, but with sufficient expertise and insight, can now analyse the same data, supporting, modifying and maybe even contradicting the results of the original study. This is the true spirit of peer review. What is more, there is a growing number of initiatives by multiple journals to make code and intermediate data available as well (Hrynaszkiewicz, 2017), making a bioinformatic study completely transparent, offering the experience inherent in its code as a resource and a learning tool for researchers working on similar projects. Just as importantly, it allows for the replicability of the study, ensuring scientific integrity and quality of results.

In chapter 2, we support the emerging principles of biological data science in our analysis of the most comprehensive *Lactobacillus* dataset to this day. We sequenced 175 *Lactobacillus* species as well as 26 additional genomes from eight associated genera, depositing both raw reads on SRA and assembled contigs on GenBank. The power of public data sharing was quickly demonstrated when Zheng *et al* downloaded our deposited sequence data, conducted a subset of analyses that overlapped with ours and had their study published online on September 22th 2015 (Zheng et al., 2015), a full seven days before our more comprehensive study became available.

Chapter 1 of this thesis described the historical confusion and continuing difficulty of defining the *Lactobacillus* genus, accrediting much of this to the contradiction between classifications in early phenotypic and subsequent genotypic properties. It was for this reason, and because the genus is such an important resource in industrial food fermentation and probiotics, that we sequenced the most comprehensive *Lactobacillus* genomic dataset still available (Sun et al., 2015). Our aim was neither to saturate the available sequence diversity of individual *Lactobacillus* species nor to capture the complete functional repertoire of the *Lactobacillus* genus. Our primary goals were to provide a reliable phylogenomic template for future studies, offering a dependable structure for which present and future analyses of *Lactobacillus* genomics could be fastened, and to describe the considerable functional diversity displayed by the type strains of, at that time, almost all of the characterised *Lactobacillus* species, including genera that branch within their phylogeny such as *Pediococcus* and *Weissella* as well as multiple reclassified taxa such as *Kandleria* and *Atopobium*. Another reason for focussing on the analysis of type strains is to allow previous taxonomy, largely phenotypically defined, to be compared with phylogeny.

We showed that *Lactobacillus* is more diverse than a typical family according to Average Nucleotide Identity (ANI) and Total Nucleotide Identity (TNI). This corroborates multiple studies ranging from Claesson *et al* who suggested that *Lactobacillus* taxonomy was in need of revision (Claesson et al., 2007) to Goldstein *et al* who suggested that the taxonomic complexity of *Lactobacillus* was the reason for the poor delineation of its species' antimicrobial susceptibilities (Goldstein et al., 2015). Chapter 2 strengthens the conclusions of previous studies that the paraphyletic nature of the genus reflects its outdated phenotypic classification as a coherent taxon. It echoes the conclusions of Salvetti *et al* (in prep) that the proposed *Lactobacillus* genus complex be more appropriately thought of as a lineage of multiple genera, historically grouped together by phenotypic traits like carbohydrate metabolism and lactic acid production rather than by the evolutionary distance inferred from core genes.

Chapter 2 also highlights the diversity of a number of important functional groups including glycosyl hydrolases, sortases, cell-envelope proteases and CRISPR-cas genes. This level of functional diversity is not actually that surprising in *Lactobacillus*, given the degree of phylogenomic diversity and horizontal gene

transfer that has also been described (as reviewed in chapter 1). What is most important is not the demonstration of diversity itself, but the detailed catalogue of functional variation that such a dataset enables. To draw a comparison, it is easy to say that human beings vary in their genetics; this has no potential whatsoever. But to characterise in detail the individual and ethnic single nucleotide variants that are persistent in our populations means that we can potentially treat and cure a range of human diseases (Enriquez and Gullans, 2015). The same is true for *Lactobacillus*; what it means for biotechnology, what it means for human health and what it means for our expanding knowledge of the ecology and evolutionary biology of these intriguingly diverse microbes.

We recognised early on that a project striving to describe both the phylogeny of such a complex group of microbes and to characterise its functional diversity would need an international team of researchers in order to optimise the study. The nature of this collaboration is summarised in the author list, from research labs around Ireland to Italy, America and China, all applying their expertise toward one study, eager to contribute and be part of a larger enterprise. As a result, we published a very novel study of the first comprehensive *Lactobacillus* phylogenomic dataset in *Nature Communications* in 2015.

Chapter 3 builds on the achievements of chapter 2 in several respects. It focuses on the functional diversity of a single species, *Lactobacillus salivarius*, removing much of the phylogenomic complexity of chapter 2 while emphasising the functional variation that exists within this reportedly probiotic species, both on the chromosomal level and on the much more variable genomic regions of its mega and smaller plasmids (Harris et al., 2017). As part of this study, we sequenced 29 strains of *L. salivarius*, depositing the genomes online in order to complement the 13 genomes then available.

For chapter 3, we adopted an even more data-centric approach, choosing to submit to *Microbial Genomics*, a new journal with a growing reputation that promotes double-blind peer review and an open data policy where all data and all code not suitable for Methods must be available at the time of submission. We provided digital online identifiers (DOI) to figshare for six in-house scripts that were coded as part of our study, making them available at the date of publication. The nature of collaboration in chapter 3 was longitudinal rather than contemporary, building on research conducted in our lab over more than ten years, adding a strong

bioinformatic component to functional and phylogenetic results from previous studies.

Chapter 4 returns to the *Lactobacillus* genus, exploring the evolutionary rates and varying selection pressures acting on an expanded version of the dataset from chapter 2. We estimated the evolutionary rates of every homologous codon for hundreds and thousands of sequences of the *Lactobacillus* core genes, summarising all these data at the level of phylogenetic sub-clades. The fact that we could do this is a testament to the ability of bioinformatics to automate computational procedures on sequence data, made possible by the use of a powerful Linux server.

Although we could have included gene-specific and even intra-genic results, we chose to focus at a higher phylogenetic level, describing the average protein evolutionary rate acting over many genes for groups of species. We used one method of calculating synonymous and non-synonymous substitutions (Nei and Gojobori, 1986), and one method of correcting for multiple sequence substitutions (Holmquist et al., 1972). Perhaps an intra-genic focus is better served by multiple methods, removing possible inaccuracies due to biological assumptions inherent in one or several of the algorithms used. The inclusion of this level of methodological comparison for specific gene sequences would have reduced the coherence of the chapter so we chose to postpone these types of analyses for a later date and another study.

The future of Microbiology will be intertwined with that of Bioinformatics. As sequencing projects get more ambitious and more computationally capable, the need for biological expertise and analytical ingenuity will be even greater. This will make collaborative studies more and more necessary, bringing together scientists from different backgrounds with complementary skills and experience that could not be instilled in a single individual due to their multi-disciplinary nature. The same is true for the *Lactobacillus* genus. The newest species announcement was *Lactobacillus alii* on October 18th, isolated from scallion kimchi. The announcement of new species of *Lactobacillus* is not a rare occurrence and can only continue as the global research community expands, adding members to this already extensive group of microbes.

Public online databases of biological sequence data are a massive global resource that represent an invaluable source of information about the living world.

Access to this exponentially growing resource coupled with the necessary analytical tools is enabling us to increase our understanding of the complex ecological and evolutionary processes acting on organisms as well as the tremendous genetic diversity of life, diversity that can be exploited for human health and wellbeing.

We are no longer just making use of the existing functional diversity we find. Because of the discovery of genetic tools such as CRISPR, organisms can now be modified very precisely to behave on a biochemical level in ways for which their genomes never evolved (Zhang et al., 2014). The developing field of synthetic biology will allow us not only to harness the vast array of phenotypes encoded in the world's sequence databases, but also to add purposefully designed sequences that do not exist in nature. In the not-too-distant future, the rate of new *Lactobacillus* strains being deposited online may not be dictated so much by discovery as by the rate of genetic engineering of new strains in the laboratory. What implications will this have for *Lactobacillus* phylogeny? Will these strains ever begin to evolve outside of their laboratory conditions, as probiotics or novel starter cultures perhaps? The interplay between horizontal gene transfer and synthetically constructed genetic compounds would surely be a daunting study for any biologist.

To say that Darwin was an insightful man would be almost laughable. The genius and the determination that he applied to his single-minded exploration of the processes that shaped all species, living and extinct, is unparalleled. He probably would not be surprised by many of the insights that have emerged from biological research since his day. There are also things that he likely could not have predicted. To know that all his life's writing can be stored on a modern USB stick might be a bit bewildering, for instance. However, it is our increasing ability to consciously and methodically alter the genetics of species that brings with it as much responsibility as it does power.

Synthetic biology is an unsettling prospect for many scientists and lay people alike and an educated guess is that it would have frightened someone as sensitive as Darwin, who delayed the publication of his book for many years for fear of its reactionary effect. Change, good or bad, can often lead to fear, and the combination of Bioinformatics and Synthetic Biology is leading to multiple revolutions in Microbiology with considerable promise for human health, and the surprises are unlikely to stop any time soon. It is truly an exciting time to be a biologist.

BIBLIOGRAPHY

AMATO, P., JOLY, M., BESAURY, L., OUDART, A., TAIB, N., MONE, A. I., DEGUILLAUME, L., DELORT, A. M. & DEBROAS, D. 2017. Active microorganisms thrive among extremely diverse communities in cloud water. *PLoS One,* 12**,** e0182869.

CAMPANA, R., VAN HEMERT, S. & BAFFONE, W. 2017. Strain-specific probiotic properties of lactic acid bacteria and their interference with human intestinal pathogens invasion. *Gut Pathog,* 9**,** 12.

CHAIN, P., KURTZ, S., OHLEBUSCH, E. & SLEZAK, T. 2003. An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief Bioinform,* 4**,** 105-23.

CHARLES, P. & DURHAM 1952. *Alcohol, Culture and Society*, Duke University Press.

CLAESSON, M. J., JEFFERY, I. B., CONDE, S., POWER, S. E., O'CONNOR, E. M., CUSACK, S., HARRIS, H. M., COAKLEY, M., LAKSHMINARAYANAN, B., O'SULLIVAN, O., FITZGERALD, G. F., DEANE, J., O'CONNOR, M., HARNEDY, N., O'CONNOR, K., O'MAHONY, D., VAN SINDEREN, D., WALLACE, M., BRENNAN, L., STANTON, C., MARCHESI, J. R., FITZGERALD, A. P., SHANAHAN, F., HILL, C., ROSS, R. P. & O'TOOLE, P. W. 2012. Gut microbiota composition correlates with diet and health in the elderly. *Nature,* 488**,** 178-84.

CLAESSON, M. J., VAN SINDEREN, D. & O'TOOLE, P. W. 2007. The genus Lactobacillus--a genomic basis for understanding its diversity. *FEMS Microbiol Lett,* 269**,** 22-8.

DARWIN, C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life,* London, John Murray.

ENRIQUEZ, J. & GULLANS, S. 2015. *Evolving Ourselves,* London, Oneworld Publications.

FLINT, H. J., SCOTT, K. P., DUNCAN, S. H., LOUIS, P. & FORANO, E. 2012. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes,* 3**,** 289-306.

GOLDSTEIN, E. J., TYRRELL, K. L. & CITRON, D. M. 2015. Lactobacillus species: taxonomic complexity and controversial susceptibilities. *Clin Infect Dis,* 60 Suppl 2**,** S98-107.

HARRIS, H. M. B., BOURIN, M. J. B., CLAESSON, M. J. & O'TOOLE, P. W. 2017. Phylogenomics and comparative genomics of Lactobacillus salivarius, a mammalian gut commensal. *Microb Genom,* 3**,** e000115.

HOLMQUIST, R., CANTOR, C. & JUKES, T. 1972. Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids. *J Mol Biol,* 64**,** 145-61.

HOOK, S. V. 2011. *Louis Pasteur: Groundbreaking Chemist and Biologist,* Minnesota, USA, ABDO Publishing Company.

HRYNASZKIEWICZ, I. 2017. The importance of policy for enabling research data sharing. *Higher education: policy, people and politics.* [Online]. Available from: http://wonkhe.com/blogs/the-importance-of-policy-for-enabling-research-data-sharing/.

LANE, N. 2015. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci,* 370.

NEI, M. & GOJOBORI, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol,* 3**,** 418-26.

RASKO, D. A., ROSOVITZ, M. J., MYERS, G. S., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R., CHAUDHURI, R., HENDERSON, I. R., SPERANDIO, V. & RAVEL, J. 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol,* 190**,** 6881-93.

SUN, Z., HARRIS, H. M., MCCANN, A., GUO, C., ARGIMON, S., ZHANG, W., YANG, X., JEFFERY, I. B., COONEY, J. C., KAGAWA, T. F., LIU, W., SONG, Y., SALVETTI, E., WROBEL, A.,

RASINKANGAS, P., PARKHILL, J., REA, M. C., O'SULLIVAN, O., RITARI, J., DOUILLARD, F. P., PAUL ROSS, R., YANG, R., BRINER, A. E., FELIS, G. E., DE VOS, W. M., BARRANGOU, R., KLAENHAMMER, T. R., CAUFIELD, P. W., CUI, Y., ZHANG, H. & O'TOOLE, P. W. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun,* 6**,** 8322.

THOMPSON, L. R., SANDERS, J. G., MCDONALD, D., AMIR, A., LADAU, J., LOCEY, K. J., PRILL, R. J., TRIPATHI, A., GIBBONS, S. M., ACKERMANN, G., NAVAS-MOLINA, J. A., JANSSEN, S., KOPYLOVA, E., VAZQUEZ-BAEZA, Y., GONZALEZ, A., MORTON, J. T., MIRARAB, S., ZECH XU, Z., JIANG, L., HAROON, M. F., KANBAR, J., ZHU, Q., JIN SONG, S., KOSCIOLEK, T., BOKULICH, N. A., LEFLER, J., BRISLAWN, C. J., HUMPHREY, G., OWENS, S. M., HAMPTON-MARCELL, J., BERG-LYONS, D., MCKENZIE, V., FIERER, N., FUHRMAN, J. A., CLAUSET, A., STEVENS, R. L., SHADE, A., POLLARD, K. S., GOODWIN, K. D., JANSSON, J. K., GILBERT, J. A. & KNIGHT, R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*.

VIOLATTI, C. 2014. Neolithic. *Ancient History.*

WANG, Y., ZHANG, R., HE, Z., VAN NOSTRAND, J. D., ZHENG, Q., ZHOU, J. & JIAO, N. 2017. Functional Gene Diversity and Metabolic Potential of the Microbial Community in an Estuary-Shelf Environment. *Frontiers in Microbiology,* 8.

WHO. 2017. *Antibiotic resistance* [Online]. World Health Organization. Available: http://www.who.int/mediacentre/factsheets/antibiotic-resistance/en/.

YU, J., BLOM, J., GLAESER, S. P., JAENICKE, S., JUHRE, T., RUPP, O., SCHWENGERS, O., SPANIG, S. & GOESMANN, A. 2017. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. *J Biotechnol,* 261**,** 2-9.

ZHANG, F., WEN, Y. & GUO, X. 2014. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum Mol Genet,* 23**,** R40-6.

ZHENG, J., RUAN, L., SUN, M. & GÄNZLE, M. 2015. A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology. *Appl Environ Microbiol,* 81**,** 7233-43.

Acknowledgements

I have been influenced by a lot of people over the years, good and bad. I was tempted to turn this section into a version of Father Ted's acceptance speech of the Golden Cleric: And now we move on to 'liars'. Instead, I'll echo that marvellous quote by Bilbo Baggins in his farewell speech to the Shire: "I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve." But I better keep things largely positive…

Mom and Dad, thank you for everything you did right. I didn't turn out too badly now, did I? Dad, your unwavering acceptance of your very un-Kerry-like son was always a cool breath of fresh air. Mom, your eccentric nature is unfortunately hereditary. To my twin brother, Ken, thank you for being so dissimilar – it allowed me to be myself and walk my own path.

I give a very mixed thank you to everyone who helped or hindered me during my undergraduate years and during my time as a Masters student. Teachers, friends and fellow students, you were the best and worst of my time at UCC. I was, for a short time, in agreement with Al Pacino in Scent of a Woman when he shouted, "If I were the man I was five years ago, I'd take a flamethrower to this place!"

But I came back and started a research position the following year under Professor Paul O'Toole so I guess I didn't hate the place too badly. Thank you, Paul, for giving me the opportunity to work in such an amazing lab with a solid team of scientists and inspiring people. My enthusiasm for Bioinformatics really ignited when I was thrown into the deep end of collaborative, high-paced research and my future career, the details of which I have no clue, very much began here.

A thank you to Ian Jeffery and Marcus Claesson, my early mentors and co-workers, for allowing me to find my feet amid the craziness of academia, and Marcus too for his supervisory role during my PhD. There are many other staff - administrative, technical and otherwise - who have given me assistance throughout my RA and PhD, too numerous to mention (well, I'm lazy), so I'll just say thank you, you're help was very much appreciated.

A non-scientific interlude for my MMA and jiu jitsu friends: thank you for keeping me sane and grounded during difficult times and for punching me enough times in the face to keep me humble whenever my ego threatened to make an appearance.

To the few friends I have in Killarney, a big thank you for not knowing much about what the hell it is I do, but for being there nonetheless. Darren Scannell, brother of the barstool, thank you for always wishing me well and being happy for my progress, convoluted as it was.

Alcohol, you have always been there to cushion my brain from this mad world and wrap it in clouds as soft as a whisper. To all the bars in Kerry and Cork, a fond thank you. We will meet again (probably tonight).

To all my friends in the Microbiology department, thank you for your support. I have seen many people come and go over the course of six years and, well, refer back to Bilbo. A special thank you to the current gang (plus a few dearly

departed) who have shared in my ups and downs, hangovers and happiness, insights and instabilities. Adam, Angela, Anna, David, Denise, Emily, Fabien, Feargal, Guy, Maurice, Max, Mrinmoy, Sidney, Tom (alphabetical to avoid favouritism – smart or what!) and all associated girlfriends, friends and lovers: ye are a great bunch of people and my time here would have been different without your idiosyncrasies and unique mixture of diagnosable psychological issues. By different, I mean it probably would have been a lot better! No, in all honesty, what is the point of achieving anything without good people to share it with. Thank you.

Page, Silvia, Elisa Salvetti, all my international colleagues and all the people with whom I share an authorship (half of whose names I can't even pronounce): thank you for making me a better scientist.

To my wonderful girlfriend Yensi, I am happy to have met you at this stage of my life when I have just enough sense to realise the importance of a good relationship. Thank you for your support and your dinners - please don't stop feeding me now that my thesis is finished. I'm finally starting to look good naked again! Oh dear, I've said too much.

Bringing it back to boring, thank you to the Health Research Board of Ireland for funding me throughout my PhD. A little more money would have been nice, but we won't fight about it.

It's getting to that stage in Acknowledgements where I'm sick of writing and I've definitely left some people out. This thank you is for those people (who would surely read this get-out-of-jail-free 'thank you' with disdain). My only consolation is that no more than a few people will read this, and my thesis will be propping up someone's computer monitor within five years.

From the biotic to the literary, a special thank you to the word 'yak'. A truly magnificent word that was once going to be the only one with which I wrote my thesis. But there will be other projects.

A final thank you to my computer 'Owen Wilson' who passed away just days after my thesis submission, serving me well for six-and-a-half years; you are stored in my memory now like my data were in yours for so many years.

As a wise Kerry man once said, "And, and… now!"

Appendix

Appendix to Chapter II

Published as Supplementary:

Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera.

Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Paul Ross R, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW.

Nat Commun. 2015 Sep 29; 6:8322.

doi: 10.1038/ncomms9322.

PMID: 26415554

SUPPLEMENTARY FIGURE LEGENDS

Supp. Figure 1. Histograms of genome size distribution (A) and GC% (B) for 175 *Lactobacillus* genomes sequenced.

Supp. Figure 2. Sizes of the pan-genome (top) and core (bottom) genomes in all 213 genomes (red) and in all genomes with less than 20, 50, 100, 200, 300, 400, 500 contigs (green).

Supp. Figure 3. Analysis of genome assembly size as a function of niche. Niche categories are plotted on the x-axis and genome assembly size in kilo base pairs is plotted on the y-axis. Box-plots represent a five-point summary of the data in the following order (from bottom to top); minimum, first quartile, median, third quartile and maximum. Outliers are represented as individual points above or below the boxplot.

Supp. Figure 4. Frequency distribution of ANI and TNI values for the *Lactobacillus* species compared to those of traditionally defined taxonomic units. The black lines indicate the frequency distribution of values for the lactobacilli, which revealed lower values for both ANI and TNI than the majority of strains within the same family but in different genera.

Supp. Figure 5. Maximum Likelihood tree. Published representative genomes that covered 452 genera from 26 phyla, as well as the 213 genomes sequenced in this research. The tree was built based on the concatenated amino acid sequences of 16 marker genes by using PhyML with 100 bootstrap iterations. The numbers at nodes are bootstrap values and the genomes included in this study were indicated by red font.

Supp. Figure 6. Maximum likelihood tree of strains of the *Lactobacillus* Genus Complex, based on the amino acid sequences of 73 core genes. The branch colors indicate different genera.

Supp. Figure 7. Distribution of glycolytic and pyruvate dehydrogenase genes across 213 lactobacilli and related species. The distribution of phosphoglycerate mutase is discriminated by the presence of genes encoding the cofactor-dependent (d) or the cofactor-independent (i)

isofunctional enzymes. For all 10 core glycolytic enzymes, gene distribution is indicated in grey-scale from absence (white) to presence of 4 gene copies (black). For the pyruvate dehydrogenase operon (4 genes), presence of a functional complex is indicated in black, and absence of a functional complex in white. The fermentation metabolism phenotype is indicated as OHO: obligately homofermentative (purple), FHE: facultatively heterofermentative (pink), and OHE: obligately heterofermentative (green).

Supp. Figure 8. Evolution of carbohydrate metabolism in the *Lactobacillus* Genus Complex. A) Maximum likelihood tree of 204 strains of the *Lactobacillus* Genus Complex based on concatenated amino acid sequence of 73 core genes. The tree was built using RAxML with 100 bootstrap iterations. B) The number of nodes and the branch lengths to the MRCA for each strain/genome. The color of the branches in panel A and the dots in panel B indicate different fermentation types, with green representing FHE, blue OHE and red OHO.

Supp. Figure 9. Heatmap illustrating the distribution and abundance of glycosyltransferase family genes across the *Lactobacillus* Genus Complex and other genera. Gene copy number of each of the 22 represented GT family members is indicated by the colour key ranging from black (absent) to green. Strains are graphed in the same order left to right as they appear top to bottom in the phylogeny (Fig. 2) with the isolation source of each strain indicated by the colour bar at the top of the heat-map.

Supp. Figure 10. Distribution of LPXTG proteins, sortases and pilus gene clusters among the 213 genomes analyzed Panel A The pilus gene clusters (PGCs) were found in 24% of all analyzed genomes and had prevalently one of the four types illustrated in Panel B. Legend: green arrow, sortase gene; blue arrow, pilin gene.

Supp. Figure 11. Comparative analysis of core CRISPR elements. The tree in panel A is derived from an alignment of the sequence of the universal Cas protein, Cas1, to create a phylogenetic tree showing the relatedness of all CRISPR-Cas systems in lactobacilli and closely related organisms (see Fig. 5A). The strain designation is followed by I, II, or III, corresponding to the respective CRISPR-Cas system type, using pink, blue and green for Type I, II and III systems, respectively. Undefined systems are represented in yellow. When multiple Cas1 proteins were found within a genome, they were differentiated by a letter. The tree in panel B is derived from an alignment of the CRISPR repeat sequences. All strain names

correlate with the master CRISPR table (Supplementary Table 6). When a strain had multiple CRISPR repeats, they were given different letters to distinguish the repeats.

Supp. Figure 12. Comparative analysis of Type II CRISPR-Cas systems. The tree in panel A is derived from an alignment of the sequence of the Type II signature Cas protein, Cas9, to create a phylogenetic tree showing the relatedness of Cas9 proteins from Type II-A and II-C systems (see Fig. 5B). The tree in panel B is derived from an alignment of the predicted tracrRNA sequences for Type II-A systems.

Supp. Figure 13. Heatmap showing the distribution of 7 phage functional categories over the 213 genomes present in the dataset. The order of columns follows the order of genomes in the phylogenetic tree in Fig. 2 from top to bottom. The colour key shows a gradation in colour from black to red to yellow to green representing gene counts from 0 to 16.

Supp. Figure 14. Heatmap and barplot showing the distribution of plasmid-associated COGs and the number of plasmids, respectively. The order of rows follows the order of genomes in the phylogenetic tree in Fig. 2 from top to bottom. The colour key shows a gradation in colour from black to red to green representing gene counts from 0 to 15.

Supp. Figure 15. Distribution and abundance of 18 different COG categories across the 213 genomes. Number of genes assigned to each of the different COG categories is indicated by the colour bar from black (absent) to green. Strains are ordered from left to right as they appear top-down in the phylogeny (Fig. 2) with source information indicated by the colour bar along the top of the heatmap.

Supp. Figure 16. Heatmap depicting the distribution and abundance of 18 insertion sequence families across the *Lactobacillus* complex and associated genera. The number of genes assigned to each IS family is indicated by the colour bar from black (absent) to green. The strains appear from left to right as they are featured top-down in the phylogeny (Fig. 2). Source information for each strain is indicated by a colour bar along the top of the heatmap.

Supp. Figure 17. Branch length distribution and TNI value distribution (1-TNI) for current phylogrouping of the *Lactobacillus* Genus complex[6], and a manually curated

phylogrouping based on the maximum likelihood tree of 73 core genes (this study; see Supplementary Figure 18).

Supp. Figure 18. Manually curated phylogrouping of the *Lactobacillus* Genus complex and associated genera based on 73 core genes maximum likelihood phylogeny. According to this revised phylogrouping, when the branch length between two strains is greater than 0.99, the probability is very high (>97.5%) that they will belong to different phylogroups, and when the branch length is less than 0.96 between two strains, the probability that they belong to the same phylogroup is > 97.5%. Compared to the existing phylogrouping[6], the adjustments made here are:

1. Two species, *L. amylotrophicus* and *L. amylophilus* that originally belonged to the *L. delbrueckii* group were excluded from *L. delbrueckii* and defined as a new Couple.
2. The single species *L. composti* was combined with the phylogroup *L. perolens*.
3. The phylogroup *L. casei* and *L. manihotivorans* were combined together with the previously defined single species, *L. camelliae*, *L. saniviri*, *L. brantae*, L. *sharpeae*, and the Couple that contained *L. thailandensis* and *L. pantheris*, was defined as a single phylogroup.
4. The single species *L.algidus* was combined with the phylogroup *L. salivarius*.
5. *Leuconostoc* and *Fructobacillus* were defined as a single phylogroup.
6. The phylogroups *L. reuteri* and *L. vaccinostercus* were combined together
7. The phylogroups *L. brevis* and *L. collinoides* and a single species, *L. malefermentans*, were combined as a single phylogroup.
8. *L. senioris* was combined with the phylogroup *L. buchneri*.
9. The couple that contained *L. ozensis* and *L. kunkeei* was combined into the phylogroup *L. fructivorans*.

Supp. Figure 19. Phylogeny inferred from a 100 core gene dataset (27 partial + 73 complete core genes).

Supp. Figure 20. Heatmap of pairwise ANI values for 213 genomes. The order of these strains is presented according to their position in the phylogenetic tree based on 73 core proteins (Fig. 2).

Supp. Figure 21. Scatterplots showing the correlation between the number of carbohydrate transport genes (y-axes) and the number of glycosyl hydrolase genes (left)

and the number of glycosyl transferase genes (right). The line of best fit for each plot was estimated using a least squares linear model.

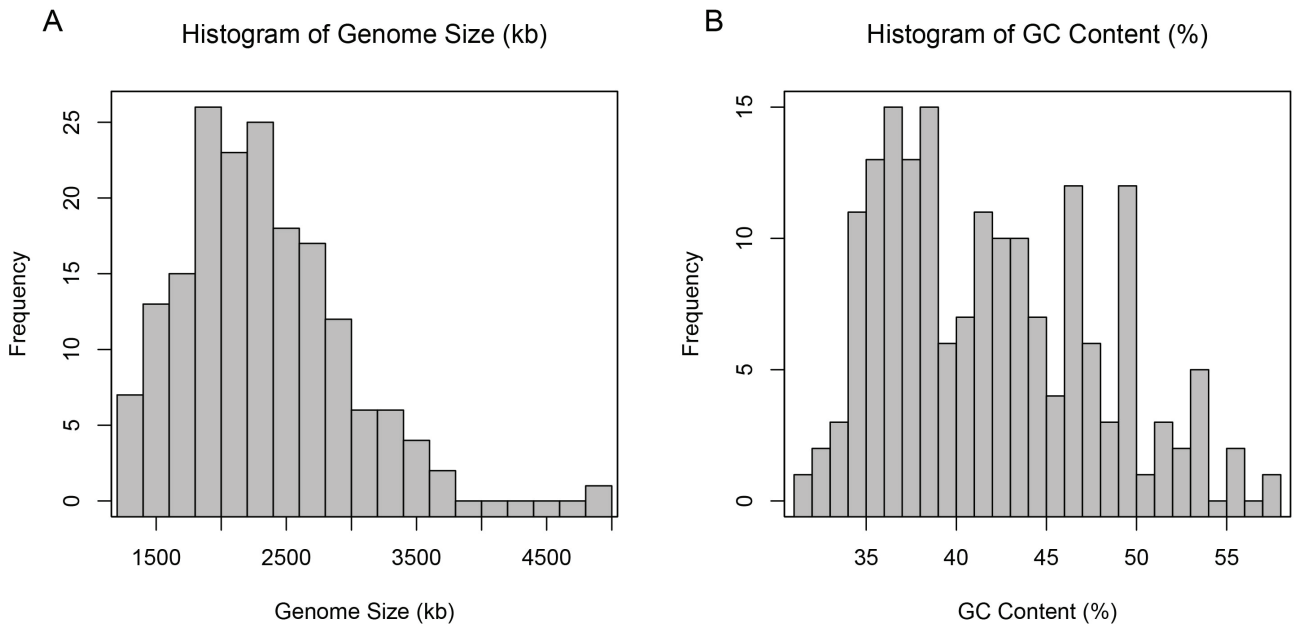
Supp. Figure 22. Barplot of the number of genes involved in carbohydrate transport for each strain. Strains are ordered according to their order in the phylogenetic tree in Fig 2.

Supp. Figure 23. The effect of normalizing counts of GHs and GTs by genome size. The three bar-graphs on the left show, from top to bottom, GH gene counts, GH gene counts expressed as a percentage of the total gene count and genome size (in kbps). The equivalent for GTs is shown in the three bar-graphs on the right. Genome size is highly correlated with total number of genes per genome (Pearson; 0.99).

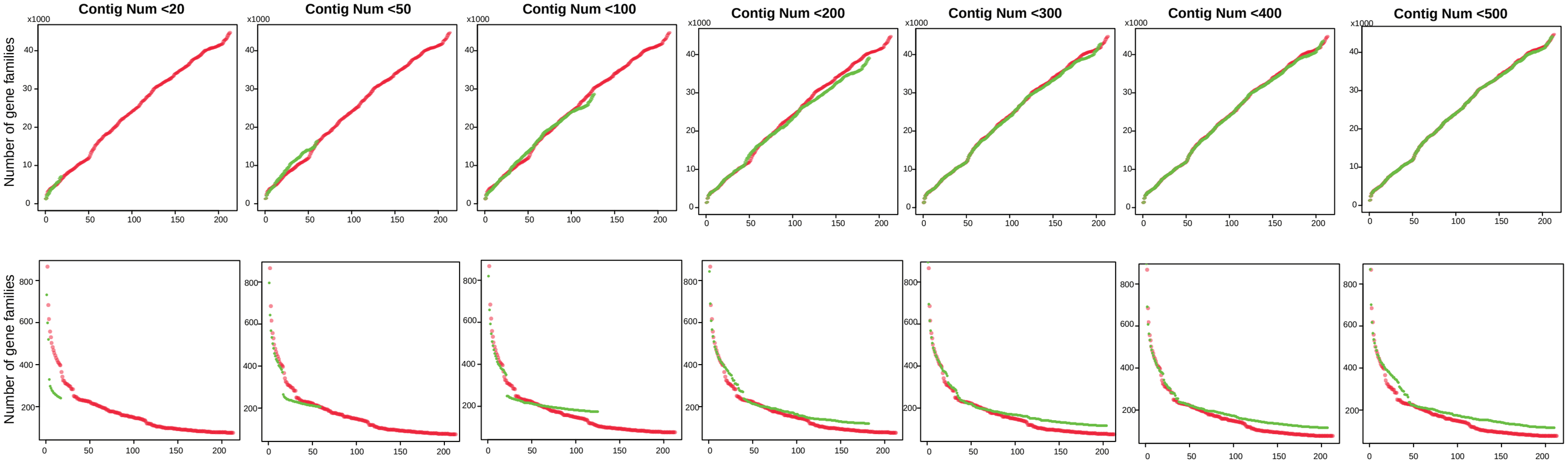
Supp. Figure 24. Heatmap illustrating the distribution and abundance of genes involved in stress response across the 213 strains. Gene copy number for 27 stress associated genes is indicated by the colour key from black (absent) to green. Type of stress each gene product confers resistance to is indicated by row names to the right of the figure. Strains are ordered from left to right as they appear top down on the phylogeny (Fig. 2) with source information for each strain indicated by a colour bar at the top of the heat-map.

Supp. Figure 25. Association of carbohydrate transport and lipid transport/metabolism with niche. Top panels display raw gene counts and bottom panels display gene counts normalized by total genes. Boxplots represent a five-point summary of the data in the following order (from bottom to top); minimum, first quartile, median, third quartile and maximum. Outliers are represented as individual points above or below the boxplot.

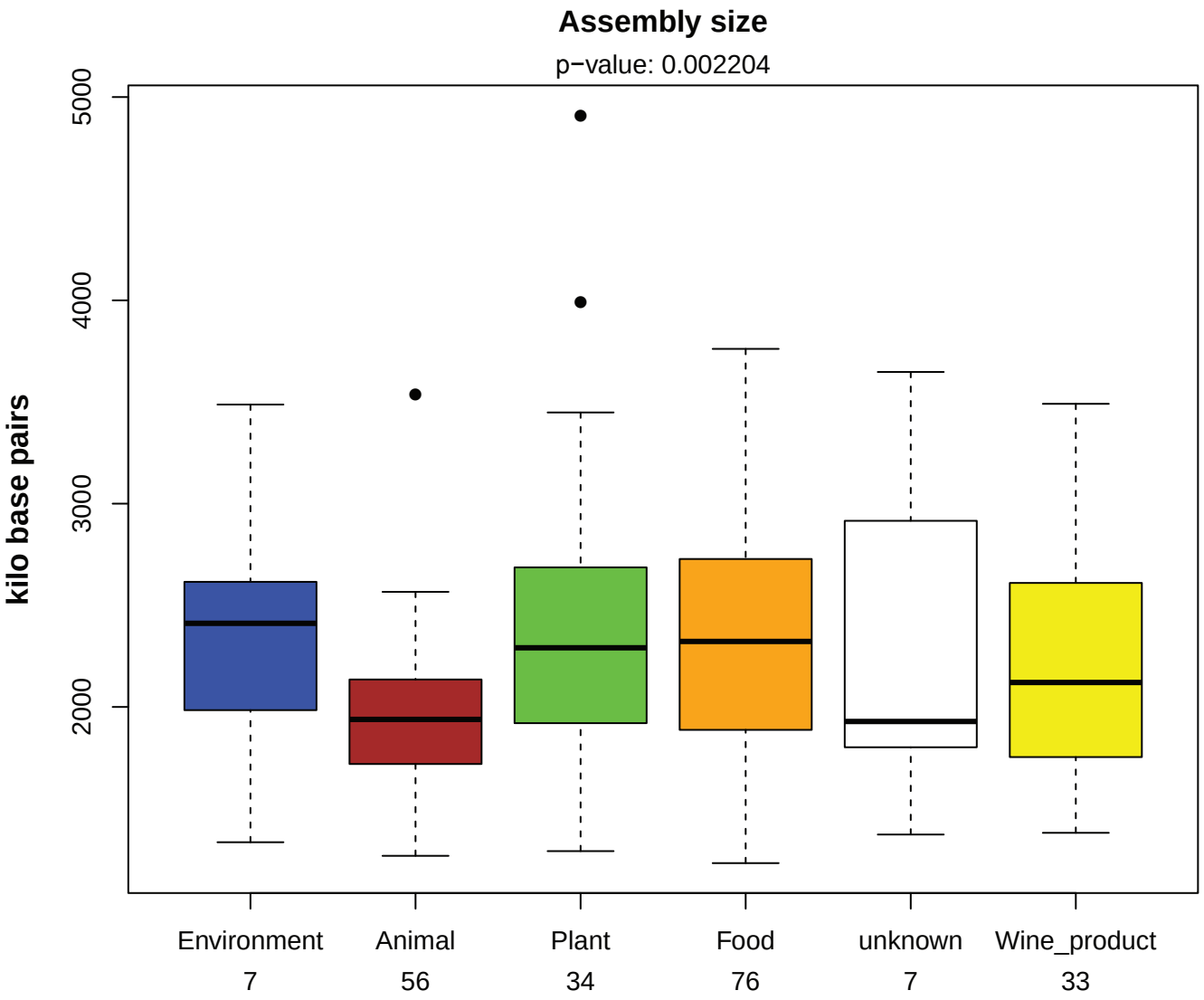
Supplementary Figures



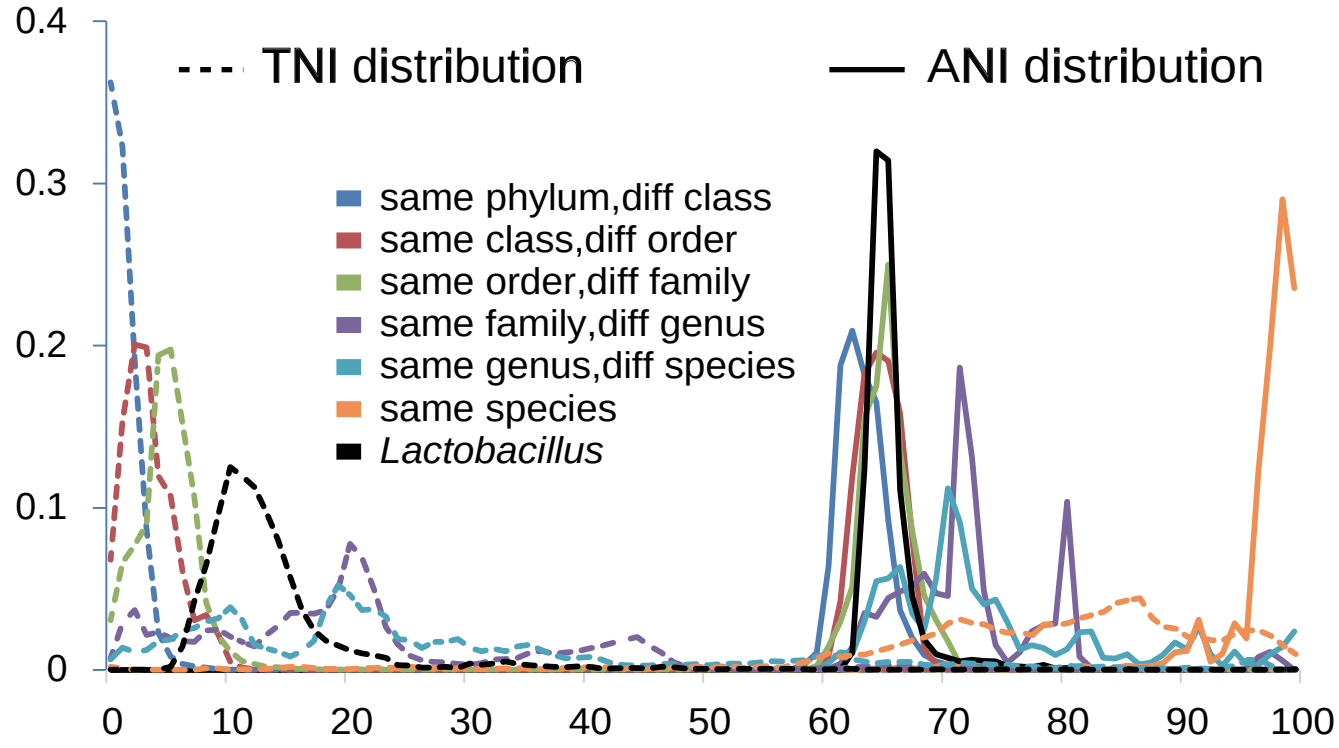
Supp. Figure 1. Histograms of genome size distribution (A) and GC% (B) for 175 *Lactobacillus* genomes.



Supp. Figure 2. Sizes of the pan-genome (top) and core-genomes (bottom) in all 213 genomes (red) and in all genomes with less than 20, 50, 100, 200, 300, 400 and 500 contigs (green).



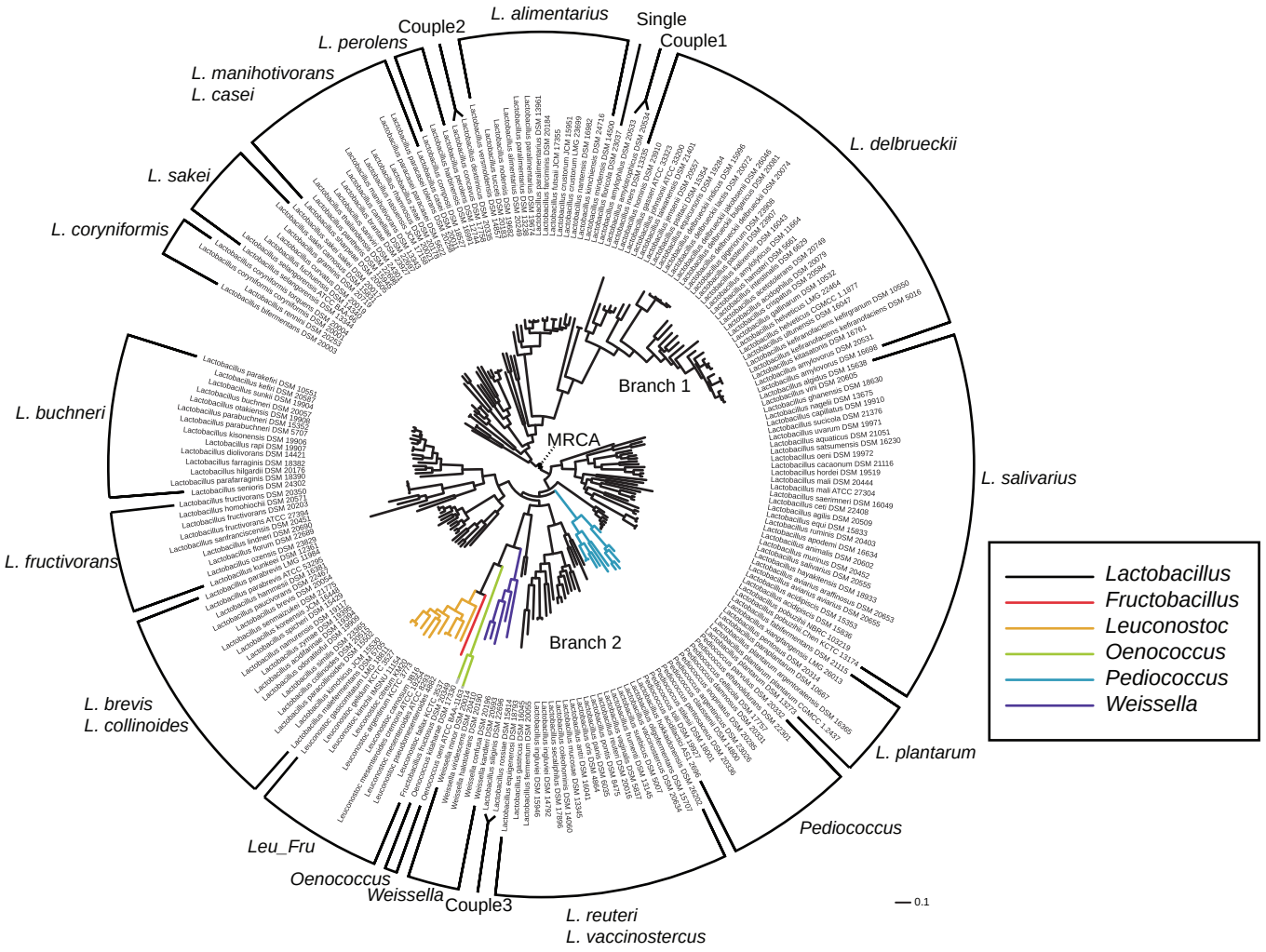
Supp. Figure 3. Analysis of genome assembly size as a function of niche. Niche categories are plotted on the x-axis and genome assembly size in kilobase pairs is plotted on the y-axis. Box-plots represent a five-point summary of the data in the following order (from bottom to top); minimum, first quartile, median, third quartile and maximum. Outliers are represented as individual points above or below the boxplot.



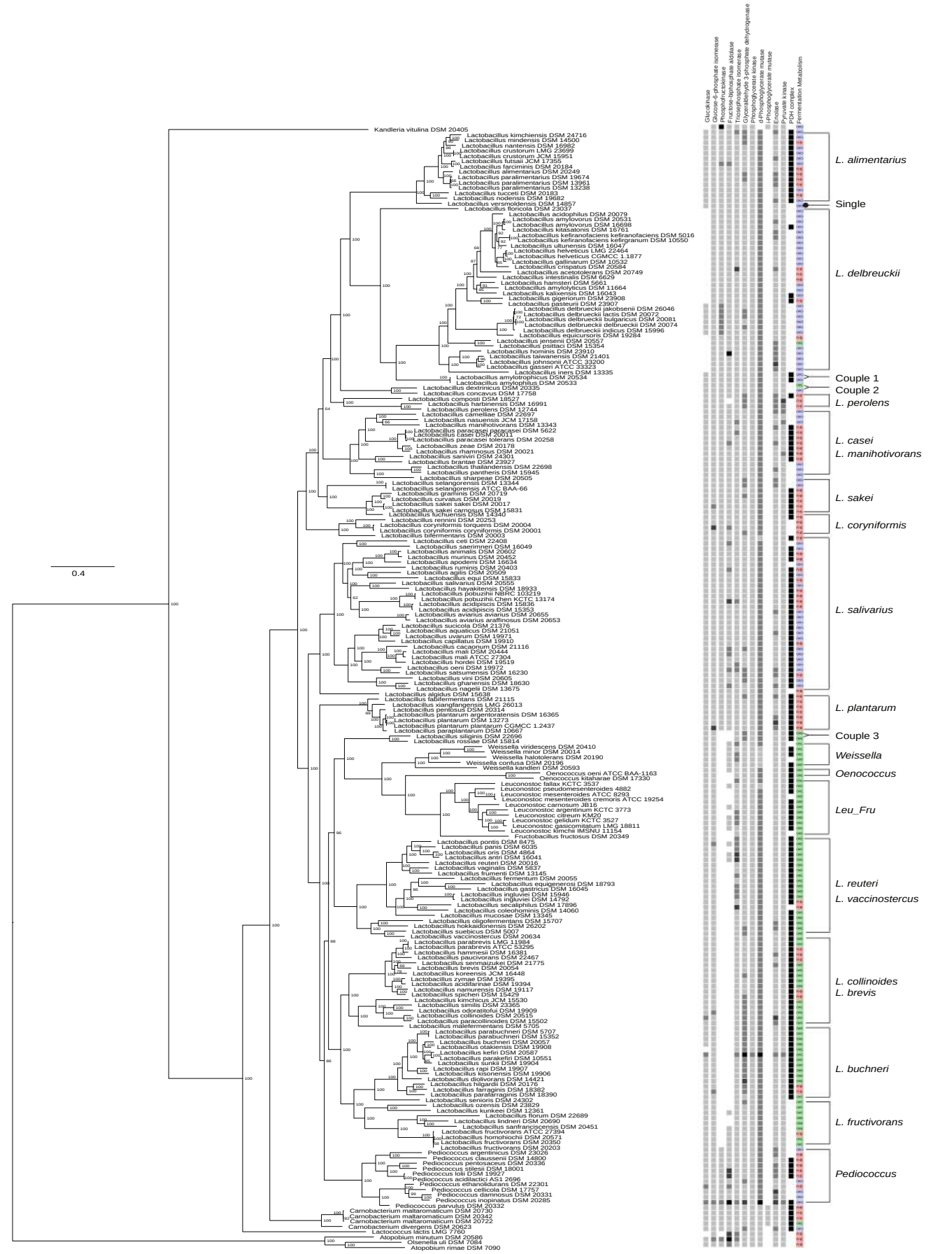
Supp. Figure 4. Frequency distribution of ANI and TNI values for the *Lactobacillus* species compared to those of traditionally defined taxonomic units. The black lines indicate the frequency distribution of values for the lactobacilli, which revealed lower values for both ANI and TNI than the majority of strains within the same family but in different genera.



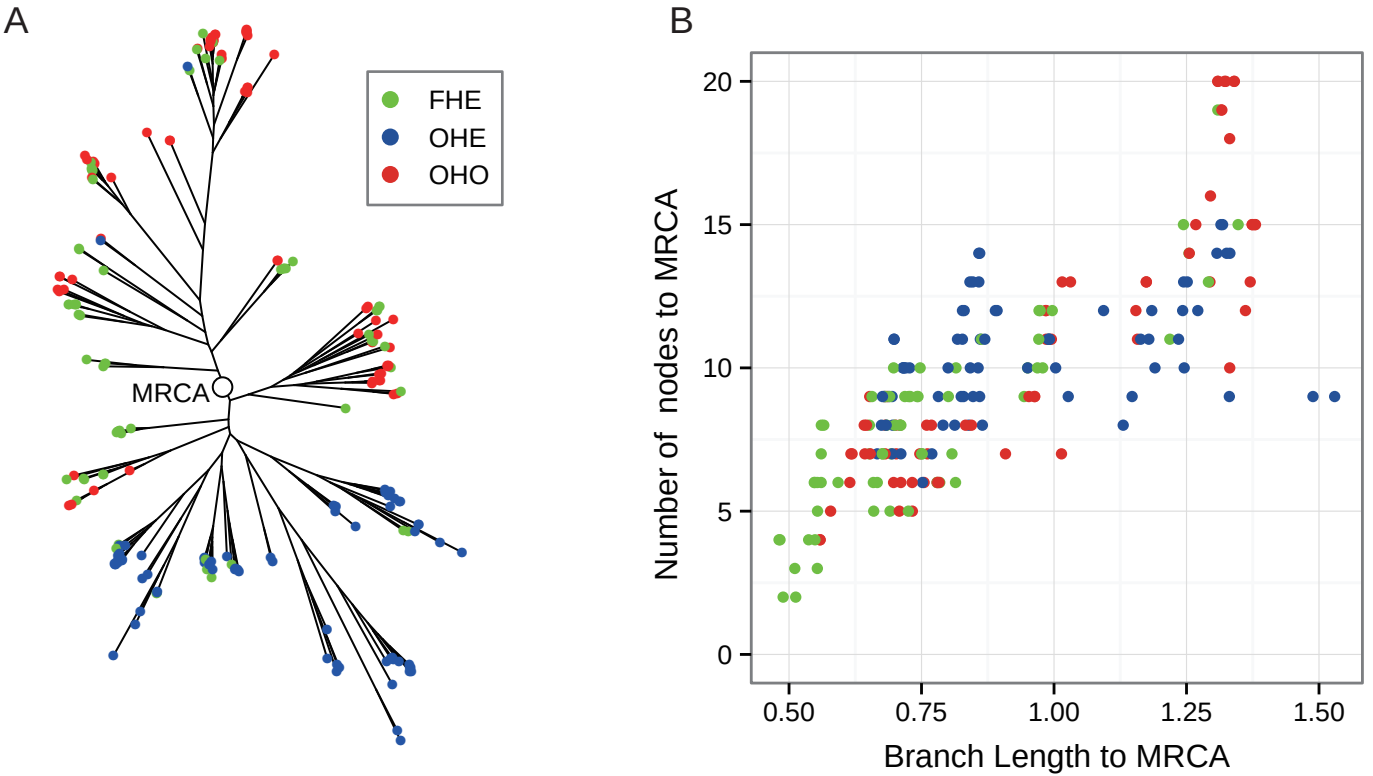
Supp. Figure 5. Maximum Likelihood tree of published representative genomes that covers 452 genera. Published representative genomes that covered 452 genera from 26 taxa, as well as the 213 genomes sequenced in this research. The tree was built based on the concatenated amino acid sequences of 16 marker genes by using RAxML with 100 bootstrap replicates. All nodes are bootstrap values and the genomes included in this study are indicated by red font.



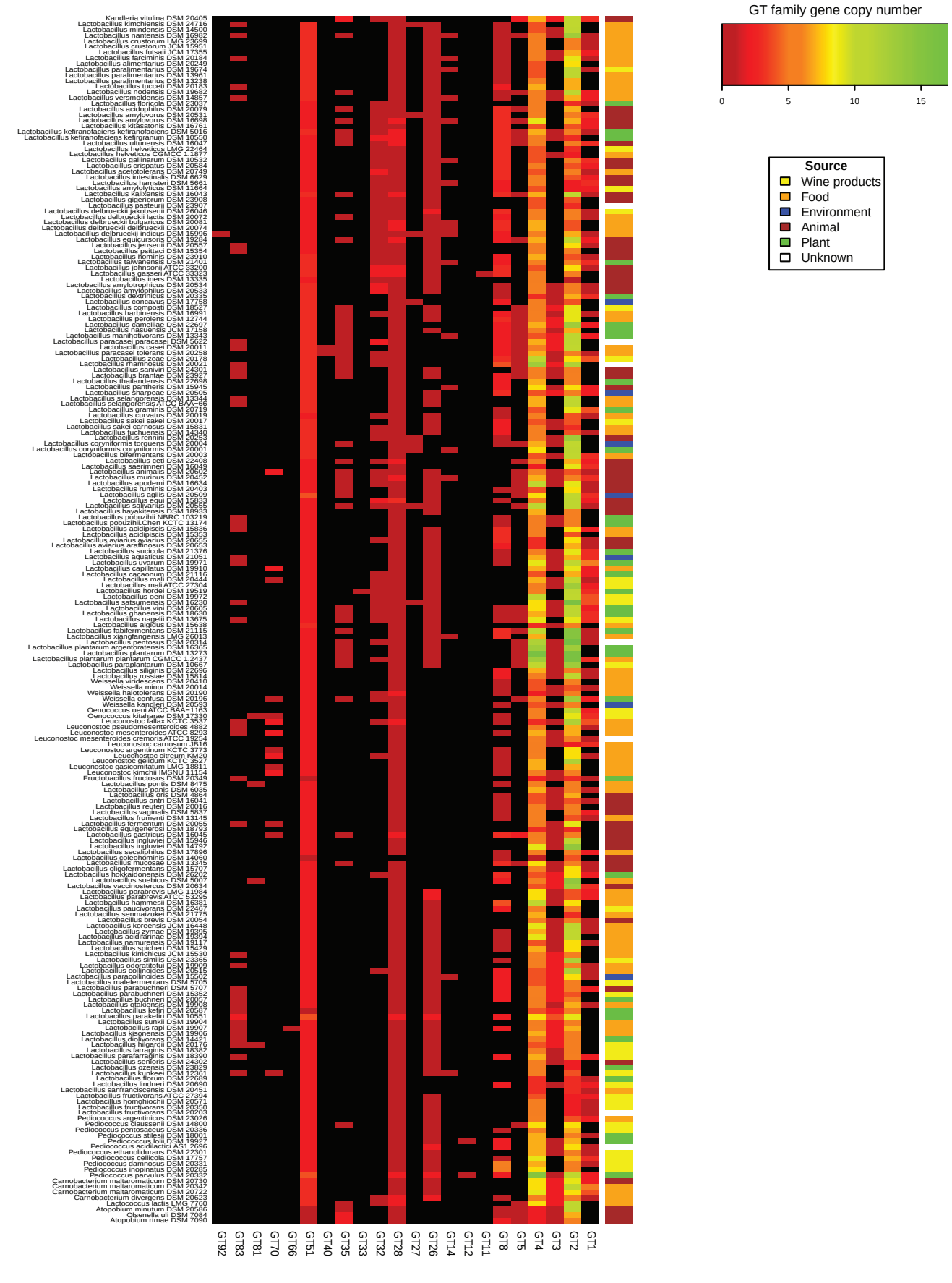
Supp. Figure 6. Maximum likelihood tree of strains of the *Lactobacillus* Genus Complex based on 73 core genes. The branch colors indicate different genera.



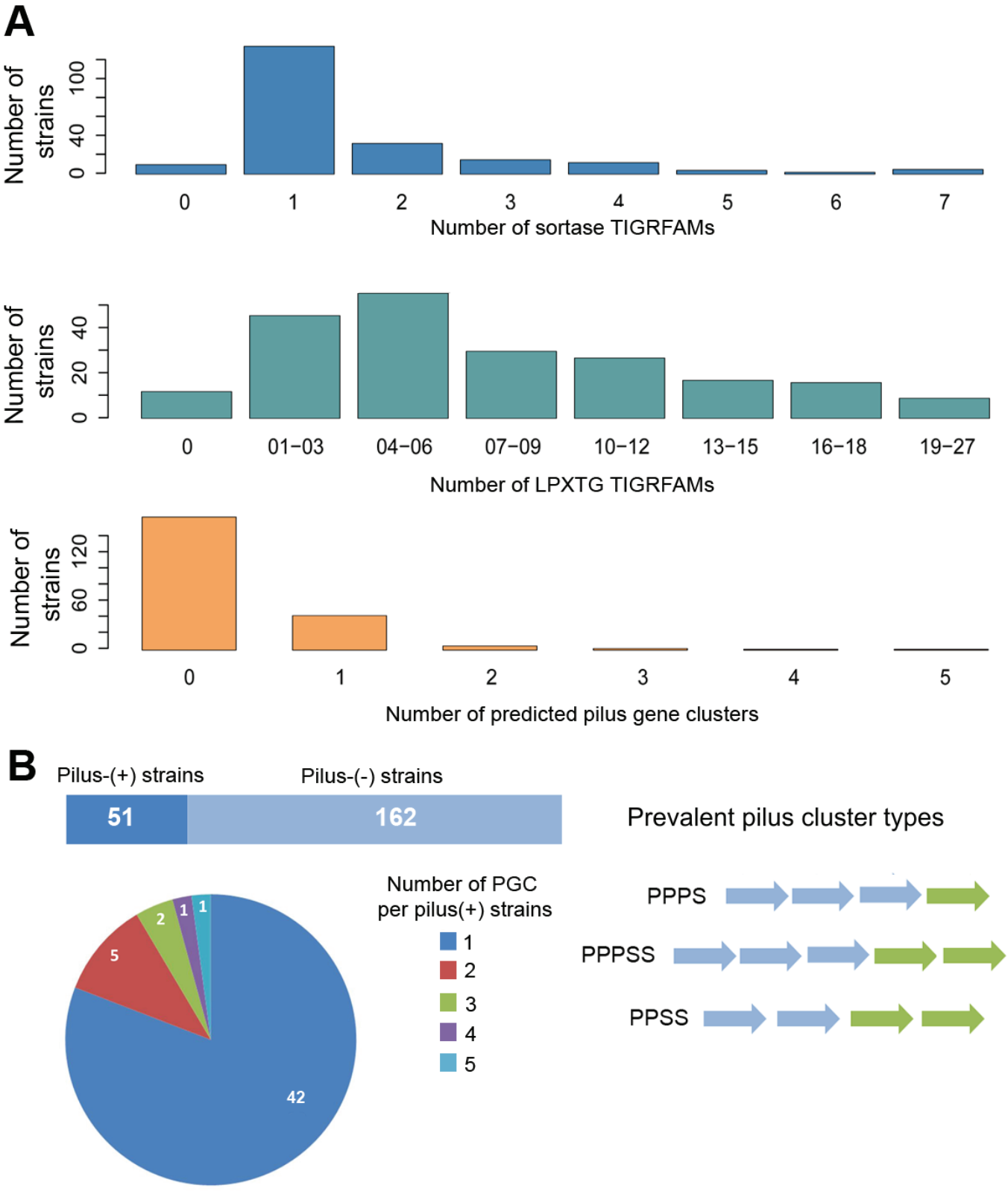
Supp. Figure 7. Distribution of glycolytic and pyruvate dehydrogenase genes across 213 lactobacilli and related species. The distribution of phosphoglycerate mutase is discriminated by the presence of genes encoding the cofactor-dependent (d) or the cofactor-independent (i) isofunctional enzymes. For all 10 core glycolytic enzymes, gene distribution is indicated in grey-scale from absence (white) to presence of 4 gene copies (black). For the pyruvate dehydrogenase operon (4 genes), presence of a functional complex is indicated in black, and absence of a functional complex in white. The fermentation metabolism phenotype is indicated as OHO: obligately homofermentative (purple), FHE: facultatively heterofermentative (pink) or OHE: obligately heterofermentative (green).



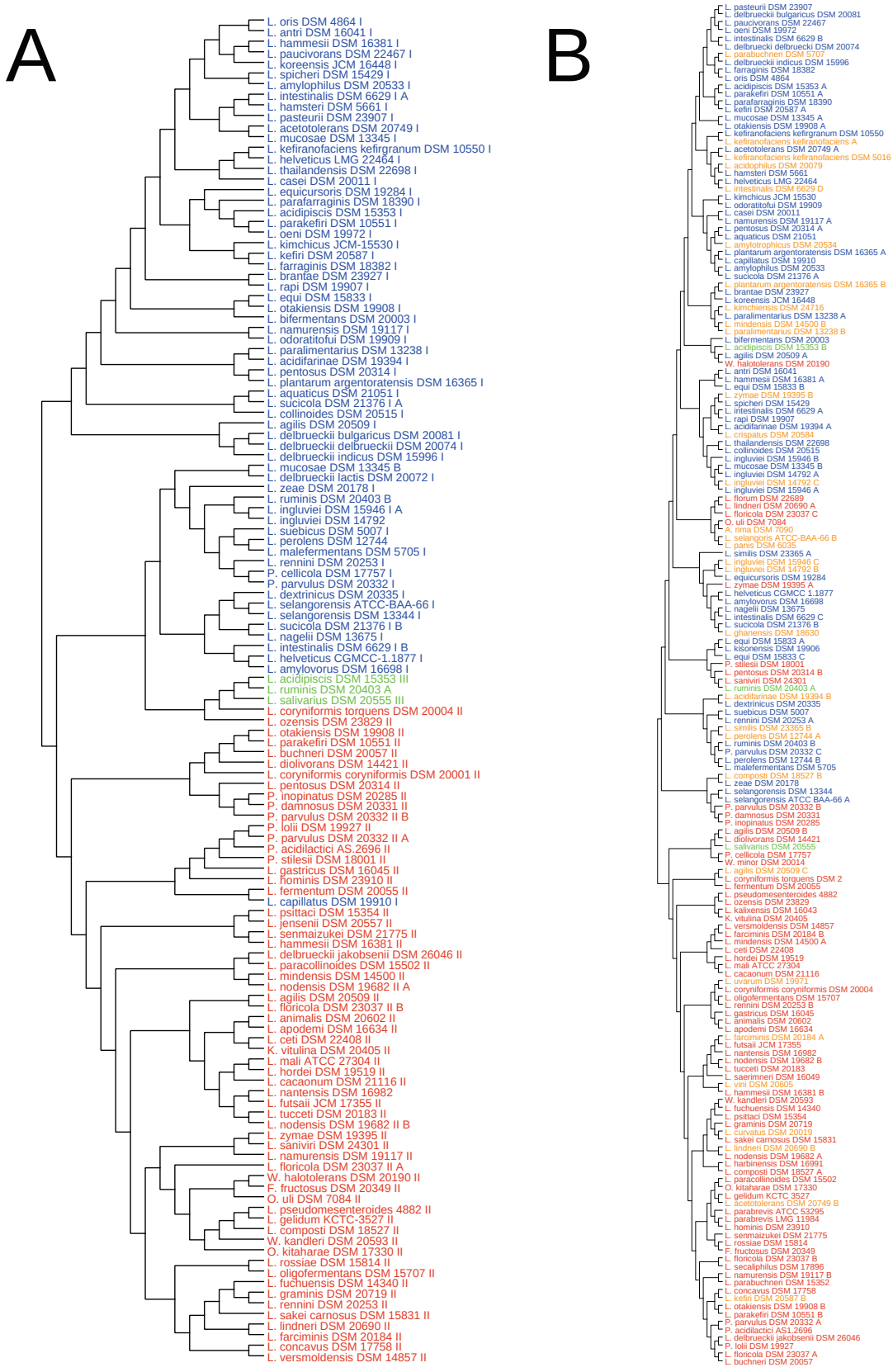
Supp. Figure 8. Evolution of carbohydrate metabolism in the *Lactobacillus* Genus Complex. A) Maximum likelihood tree of 204 strains of the *Lactobacillus* Genus Complex based on concatenated amino acid sequence of 73 core genes. The tree was built using RAxML with 100 bootstrap iterations. B) The number of nodes and the branch lengths to the MRCA for each strain/genome. The color of the branches in panel A and the dots in panel B indicate different fermentation types, with green representing FHE, blue representing OHE and red representing OHO.



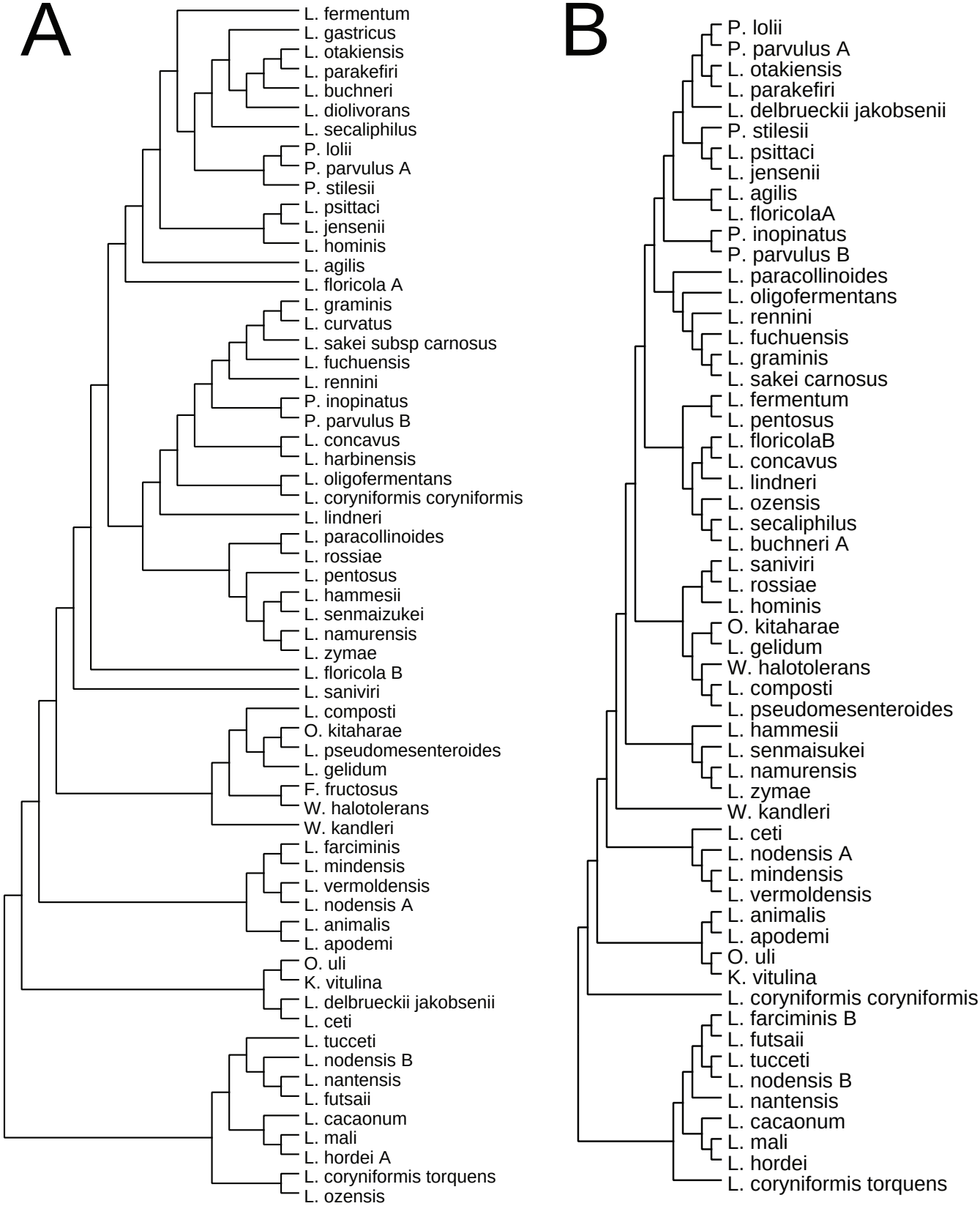
Supp. Figure 9. Heatmap of the distribution and abundance of glycosyl transferase families across the *Lactobacillus* Genus Complex and associated genera. Gene copy number of each of the 22 represented GT family members is indicated by the color key ranging from black (absent) to green. Strains are graphed in the same order from top to bottom as they appear top to bottom in the phylogeny (Fig. 2) with the isolation source of each strain indicated by the color bar at the top of the heat-map.



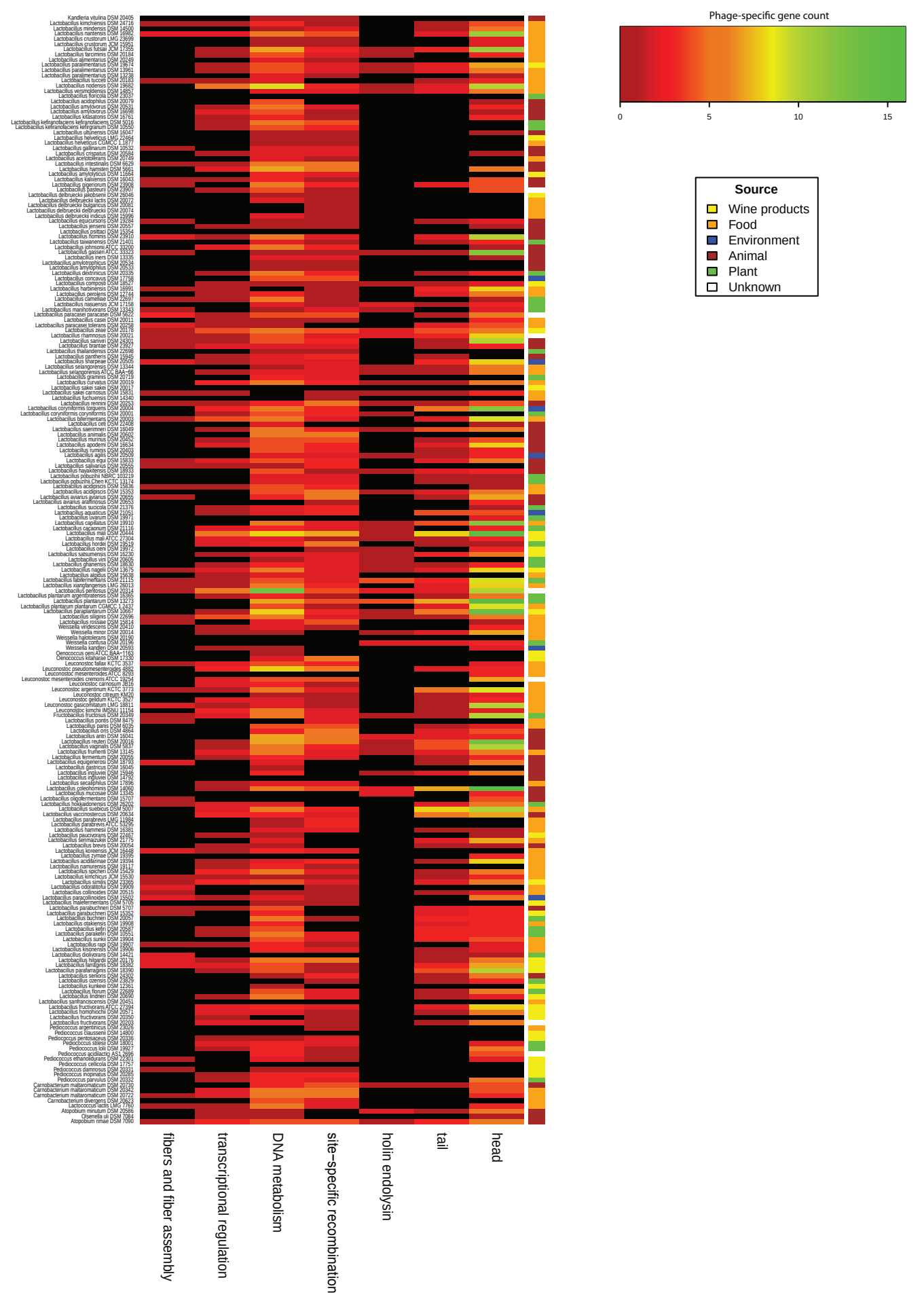
Supp. Figure 10. Distribution of LPXTG proteins, sortases and pilus gene clusters among the 213 genomes analysed. Panel A shows the pilus gene clusters (PGCs) that were found in 24% of all analyzed genomes and had prevalently one of the four types illustrated in Panel B Legend: green arrow = sortase gene; blue arrow = pilin gene.



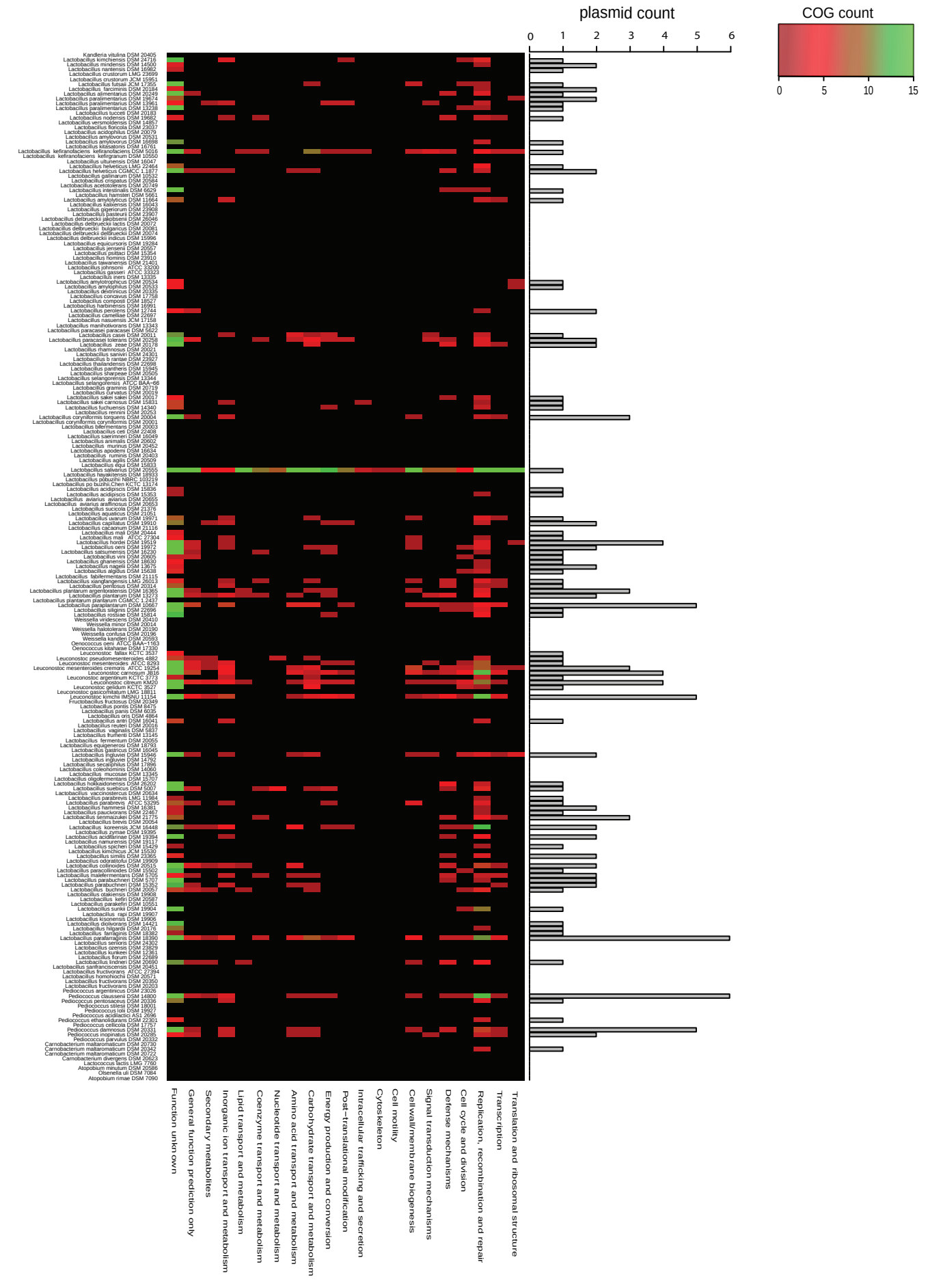
Supp. Figure 11. Comparative analysis of core CRISPR elements. The tree in panel A is derived from an alignment of the sequence of the universal Cas protein, Cas1, to create a phylogenetic tree showing the relatedness of all CRISPR-Cas systems in lactobacilli and closely related organisms (see Fig. 5A). The strain designation is followed by I, II, or III, corresponding to the respective CRISPR-Cas system type, using blue, red and green for Type I, II and III systems, respectively. Undefined systems are represented in orange. When multiple Cas1 proteins were found within a genome, they were differentiated by a letter. The tree in panel B is derived from an alignment of the CRISPR repeat sequences. All strain names correlate with the master CRISPR table (Supplementary Table 7). When a strain had multiple CRISPR repeats, they were given different letters to distinguish the repeats.



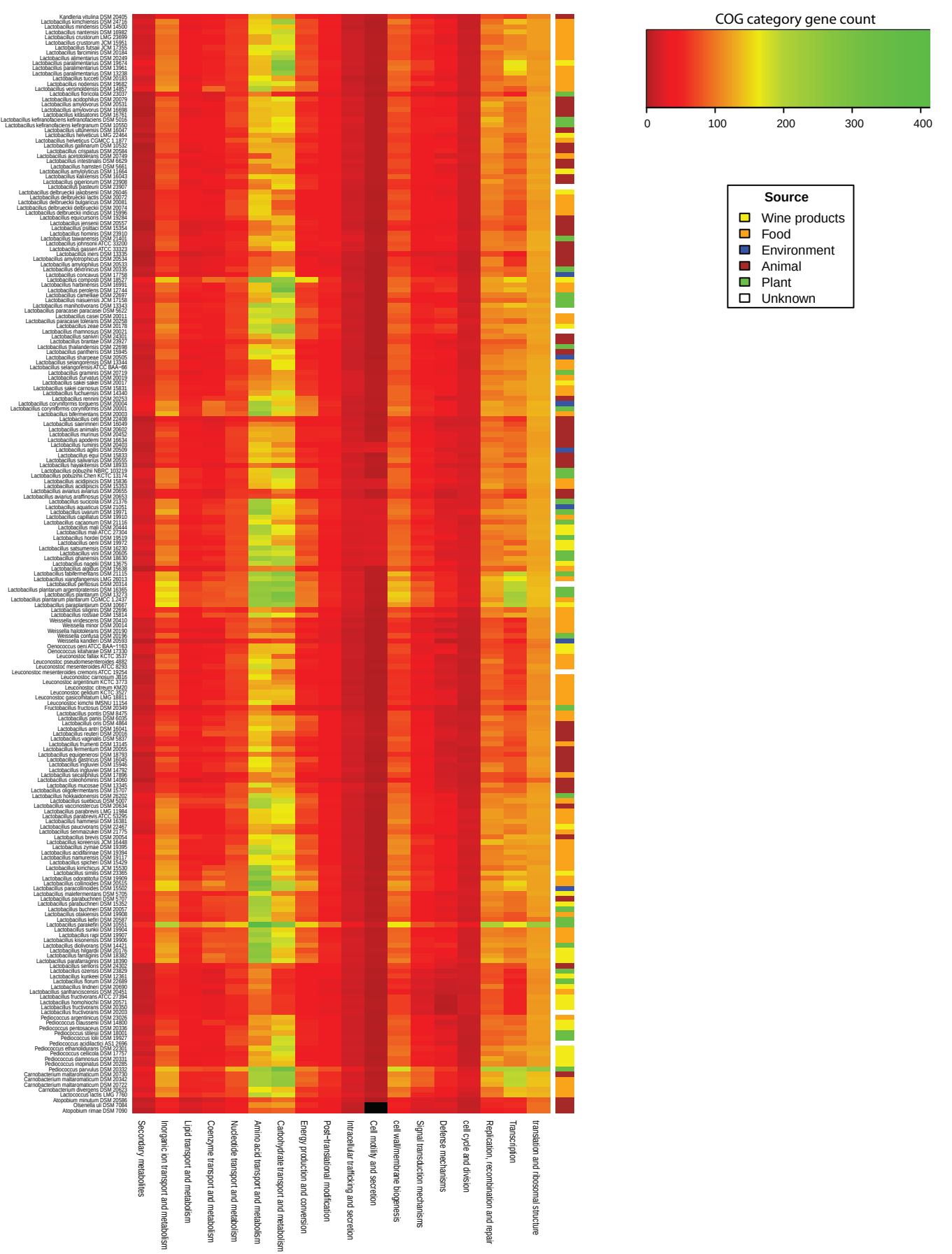
Supp. Figure 12. Comparative analysis of Type II CRISPR-Cas systems. The tree in panel A is derived from an alignment of the sequence of the Type II signature Cas protein, Cas9, to create a phylogenetic tree showing the relatedness of Cas9 proteins from Type II-A and II-C systems (see Fig. 5B). The tree in panel B is derived from an alignment of the predicted tracrRNA sequences for Type II-A systems.



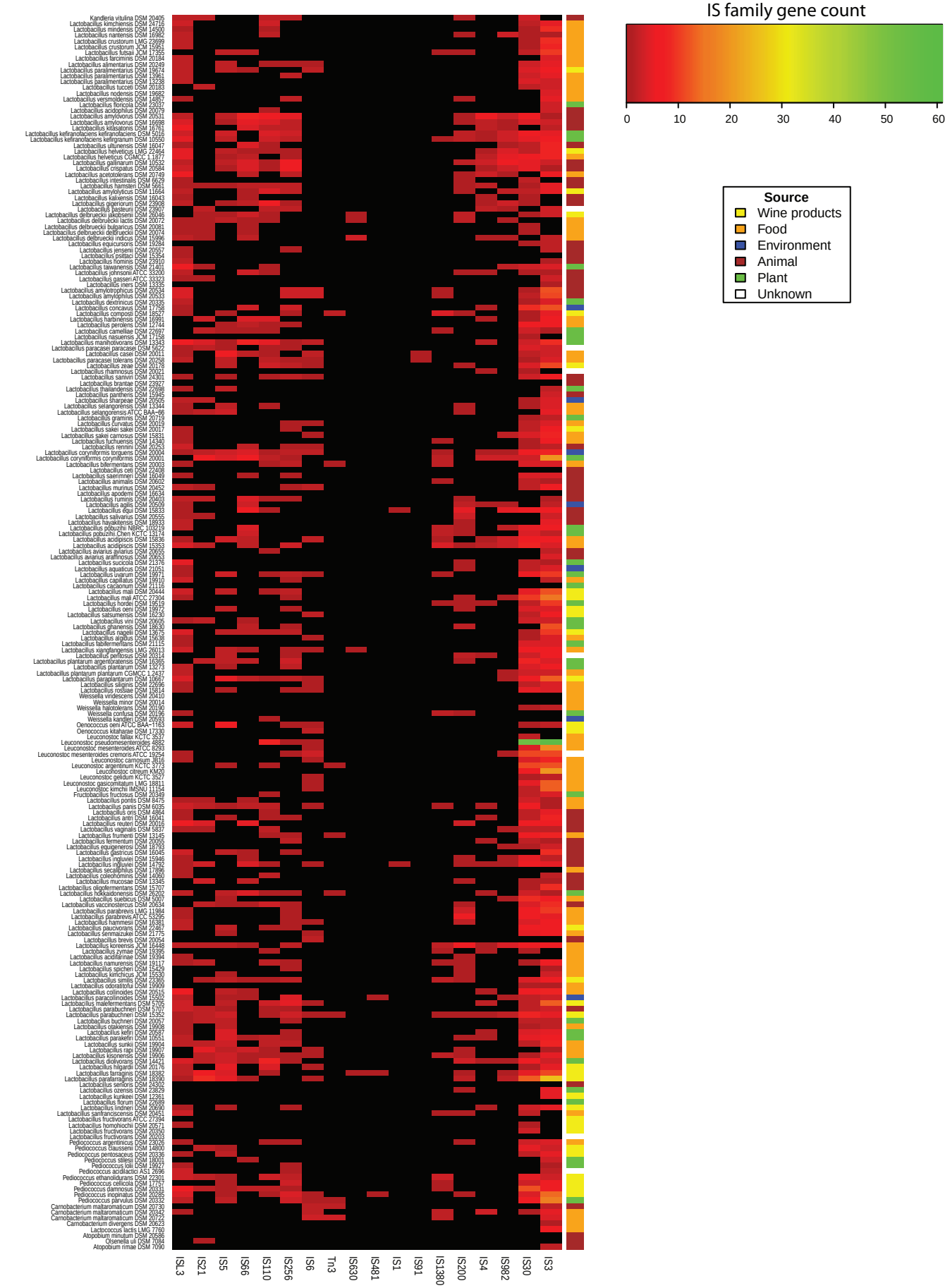
Supp. Figure 13. Heatmap of the distribution of 7 phage functional categories over the 213 genomes. The order of rows, from top to bottom, follows the order of genomes in the phylogenetic tree in Fig. 2 from top to bottom. The color key shows a gradation in color from black to red to yellow to green representing gene counts from 0 to 16.



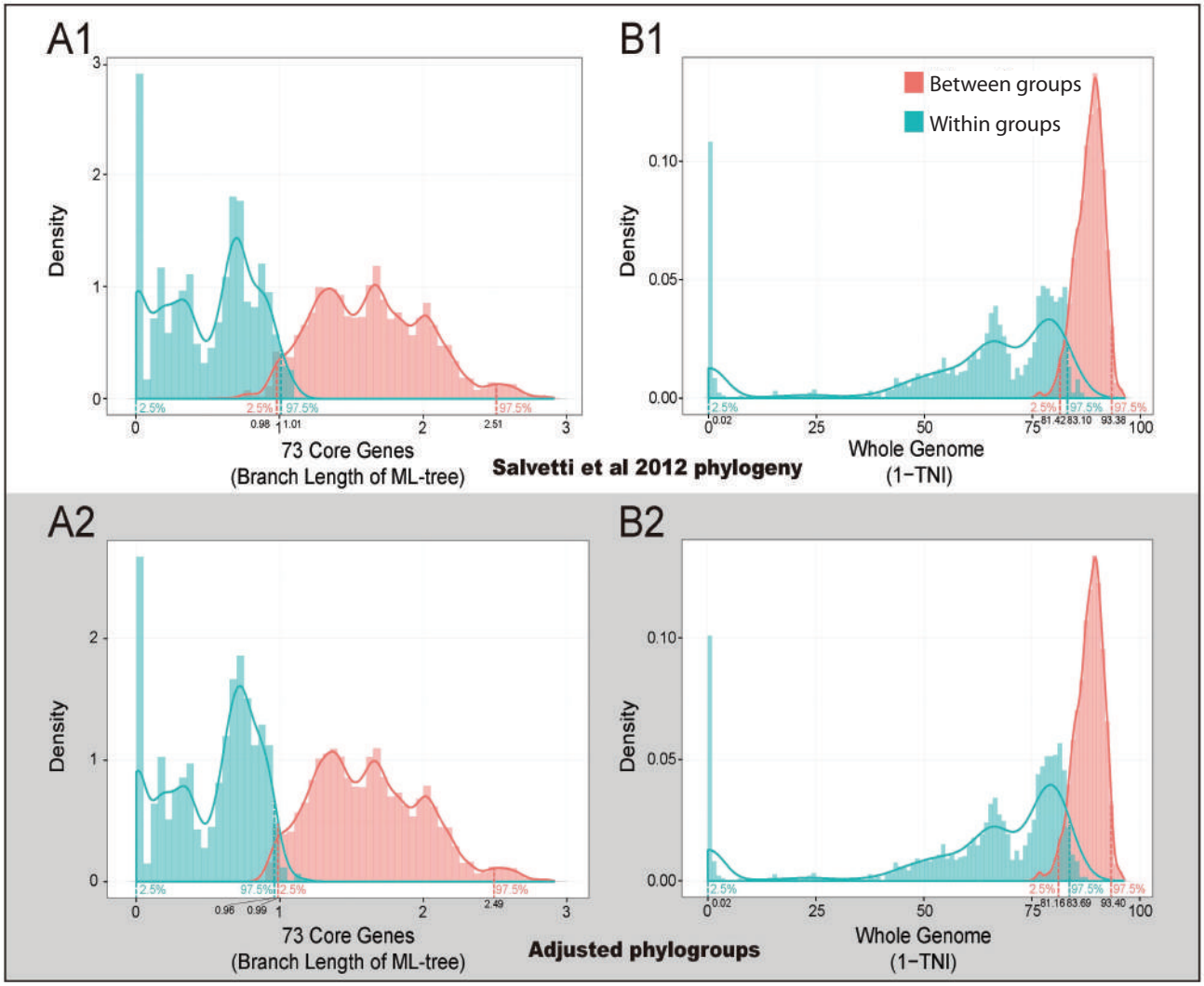
Supp. Figure 14. Heatmap and barplot showing the distribution of plasmid-associated COGs and the number of plasmids in the 213 genomes. The order of rows, from top to bottom, follows the order of genomes in the phylogenetic tree in Fig. 2 from top to bottom. The colour key shows a gradation in colour from black to red to green representing gene counts from 0 to 15.



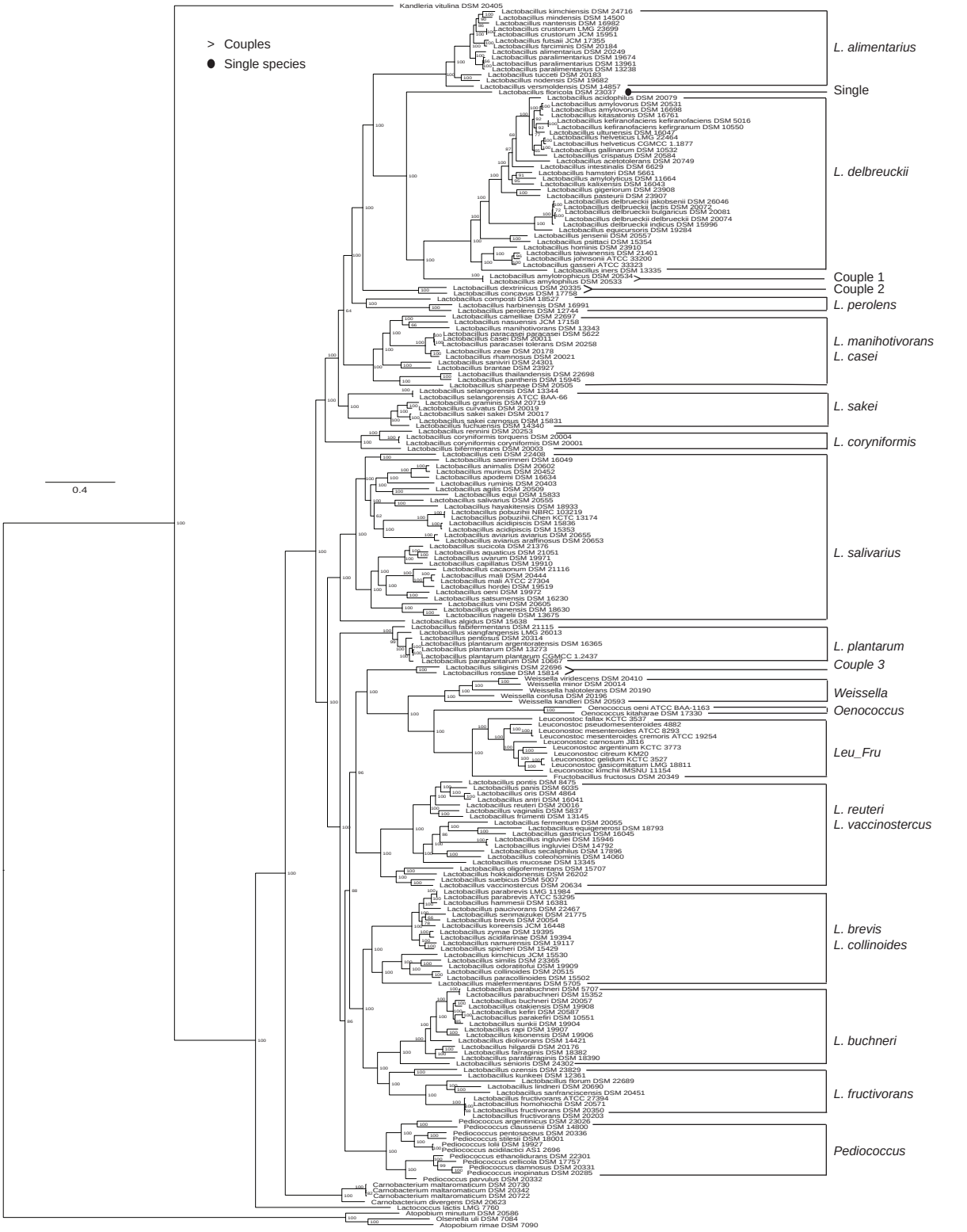
Supp. Figure 15. Heatmap of the distribution and abundance of 18 different COG categories across the 213 genomes. Number of genes assigned to each of the different COG categories is indicated by the color bar from black (absent) to green. Strains are ordered from top to bottom as they appear top-down in the phylogeny (Fig. 2) with source information indicated by the color bar along the top of the heatmap.



Supp. Figure 16. Heatmap of the distribution and abundance of 18 insertion sequence families across the 213 genomes. The number of genes assigned to each IS family is indicated by the color bar from black (absent) to green. The strains appear from top to bottom as they are featured top-down in the phylogeny (Fig. 2). Source information for each strain is indicated by a color bar along the top of the heatmap.



Supp. Figure 17. Branch length distribution and TNI value distribution (1-TNI). A current phylogrouping of the *Lactobacillus* Genus complex and a manually curated phylogrouping based on the maximum likelihood tree of 73 core genes (this study; see Supplementary Figure 18).

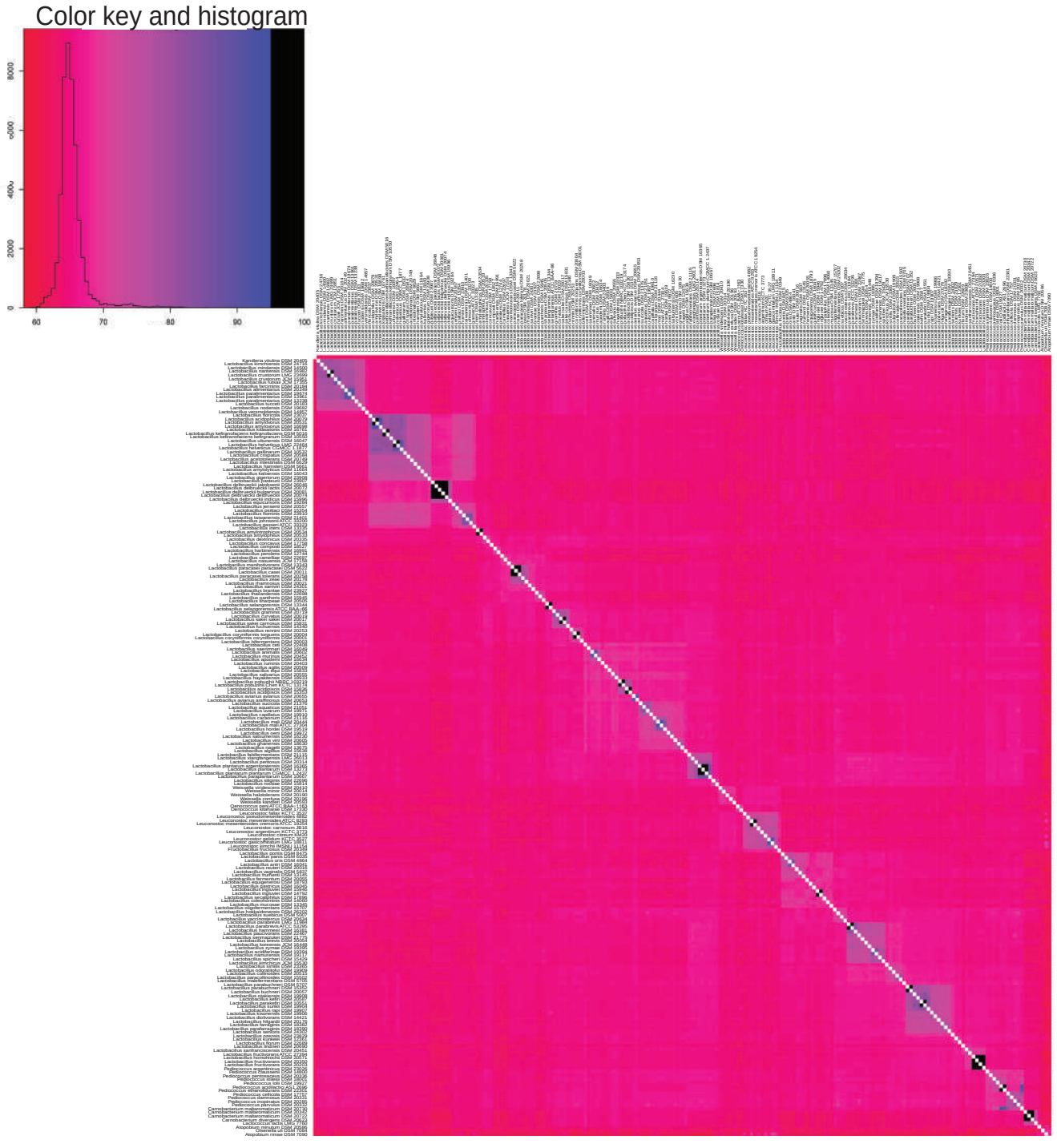


Supp. Figure 18. Manually curated phylogrouping of the *Lactobacillus* complex and associated genera based on 73 core genes maximum likelihood phylogeny. According to this revised phylogrouping, when the branch length between two strains is greater than 0.99, the probability is very high (>97.5%) that they belong to different phylogroups, and when the branch length is less than 0.96 between two strains, the probability that they belong to the same phylogroup is > 97.5%. Compared to the existing phylogrouping in Salvetti et al., 2012, the adjustments made here are:

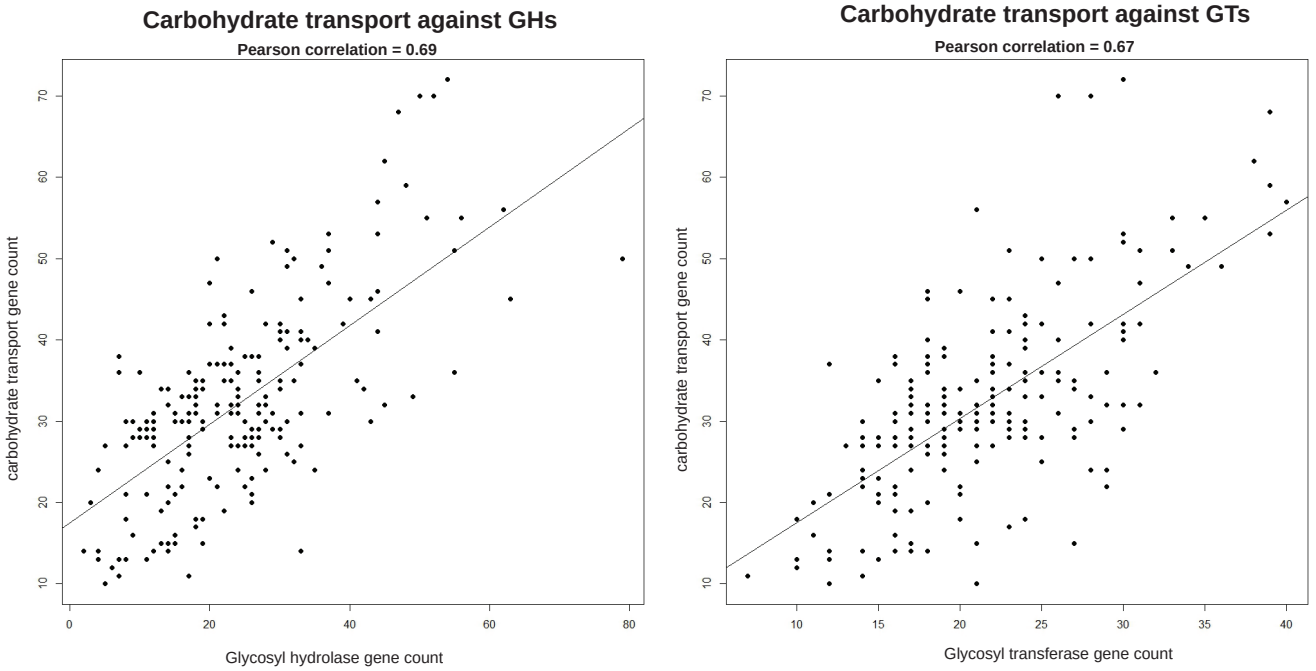
1. Two species, *L. amylotrophicus* and *L. amylophilus* that originally belonged to the *L. delbrueckii* group were excluded from *L. delbrueckii* and defined as a new Couple.
2. The single species *L. composti* was combined with the phylogroup *L. perolens.*
3. The phylogroup *L. casei* and *L. manihotivorans* were combined together with the previously defined single species, *L. camelliae, L. saniviri, L. brantae, L. sharpeae*, and the Couple that contained *L. thailandensis* and *L. pantheris*, was defined as a single phylogroup.
4. The single species *L.algidus* was combined with the phylogroup *L. salivarius.*
5. *Leuconostoc* and *Fructobacillus* were defined as a single phylogroup.
6. The phylogroups *L. reuteri* and *L. vaccinostercus* were combined together.
7. The phylogroups *L. brevis* and *L. collinoides* and a single species, *L. malefermentans,* were combined as a single phylogroup.
8. *L. senioris* was combined with the phylogroup *L. buchneri.*
9. The couple that contained *L. ozensis* and *L. kunkeei* was combined into the phylogroup *L. fructivorans.*



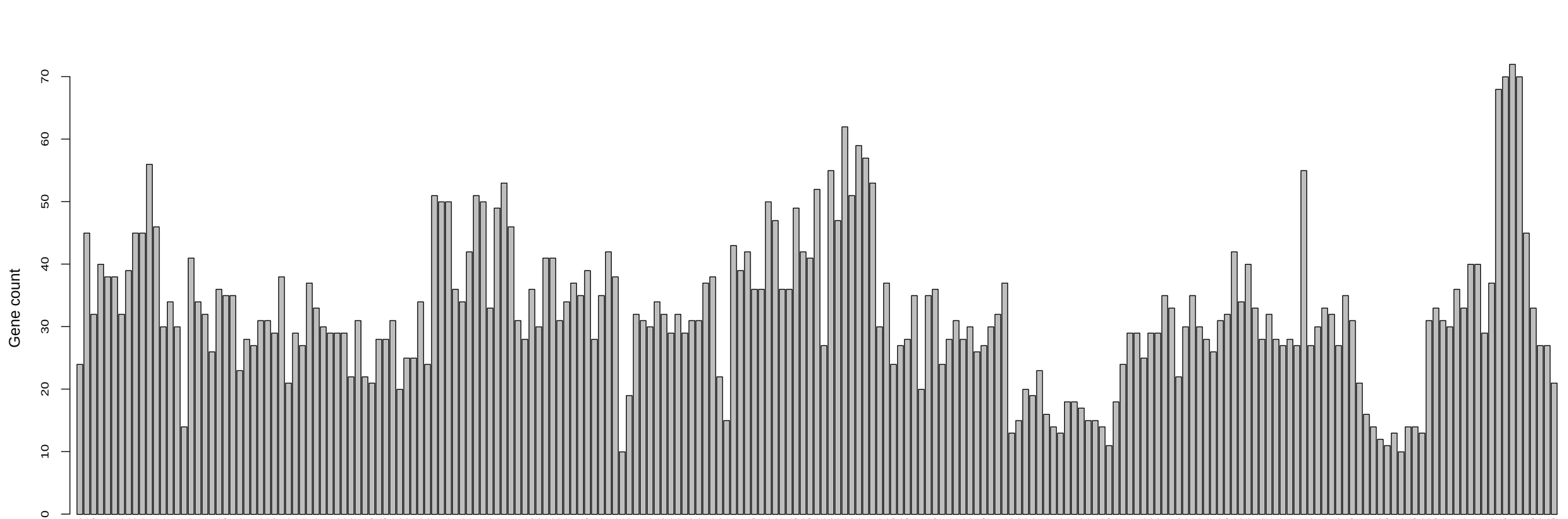
Supp. Figure 19. Phylogeny inferred from a 100 core gene dataset (27 partial + 73 complete core genes).



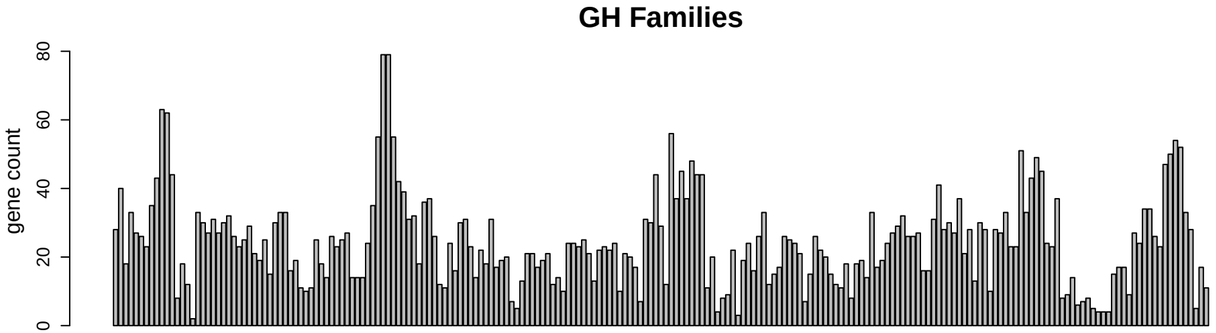
Supp. Figure 20. Heatmap of pairwise ANI values for 213 genomes. The order of the rows (top to bottom) and columns (left to right) is according to their position in the phylogenetic tree based on 73 core proteins (Fig. 2).

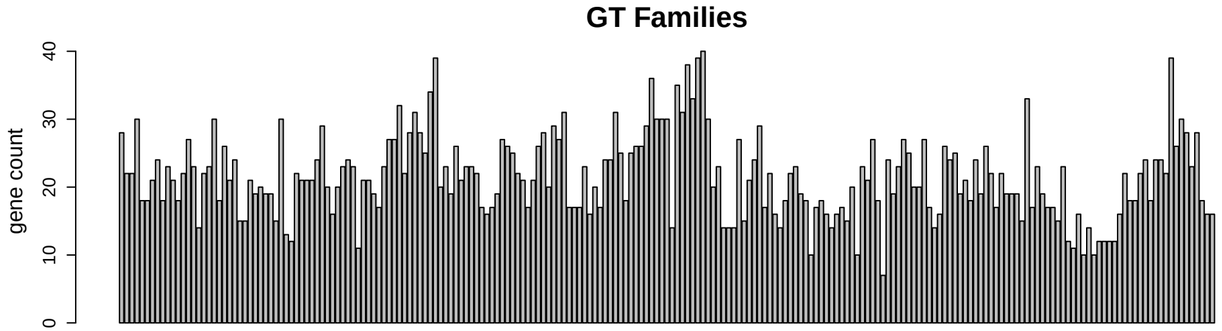


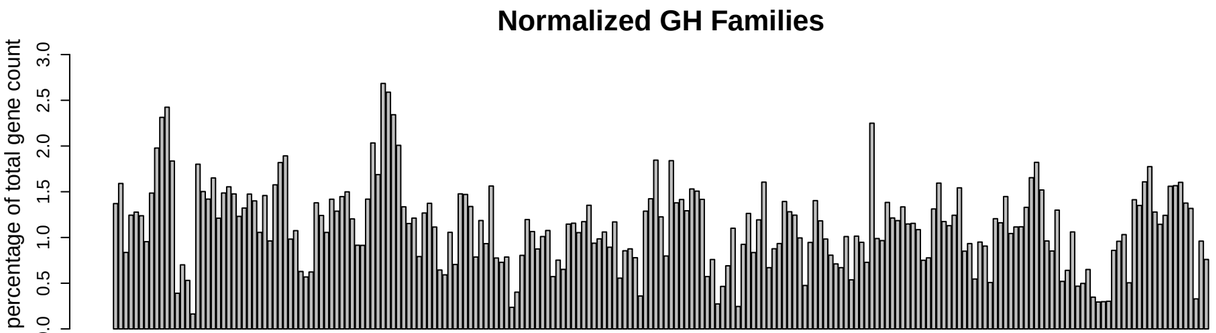
Supp. Figure 21. Scatterplots showing the correlation between the number of carbohydrate transport genes (y-axes) and the number of glycosyl hydrolase genes (x axis; left) and the number of glycosyl transferase genes (x-axis; right). The line of best fit for each plot was estimated using a least squares linear model.

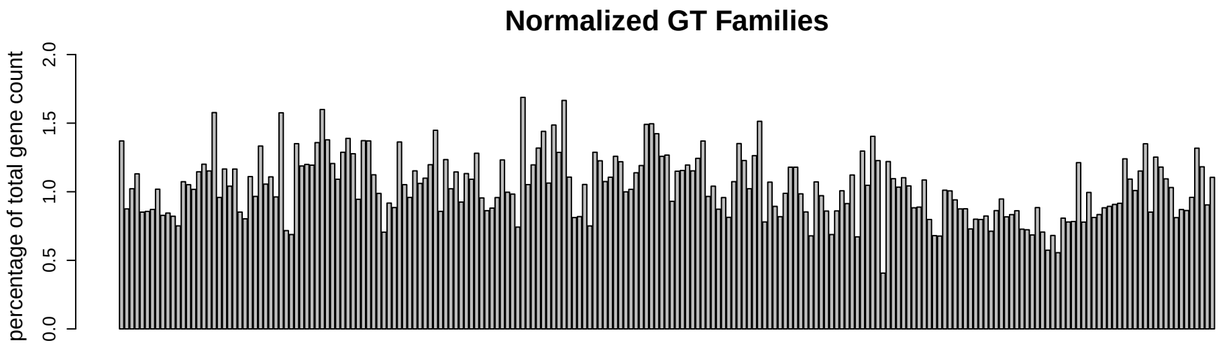


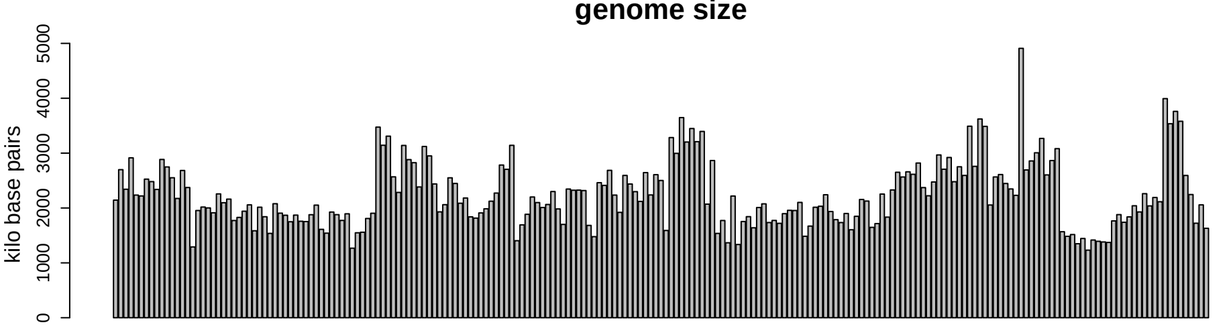
Supp. Figure 22. Barplot of the number of genes involved in carbohydrate transport for each strain. Strains are ordered according to the phylogenetic tree in Fig 2.

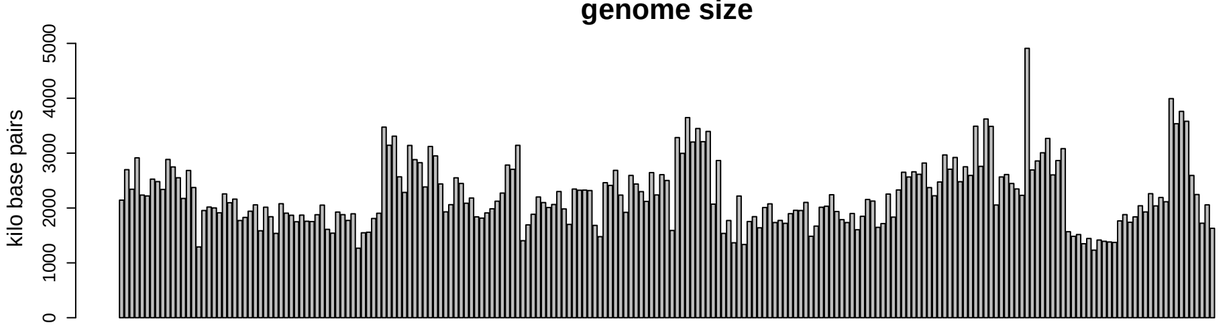




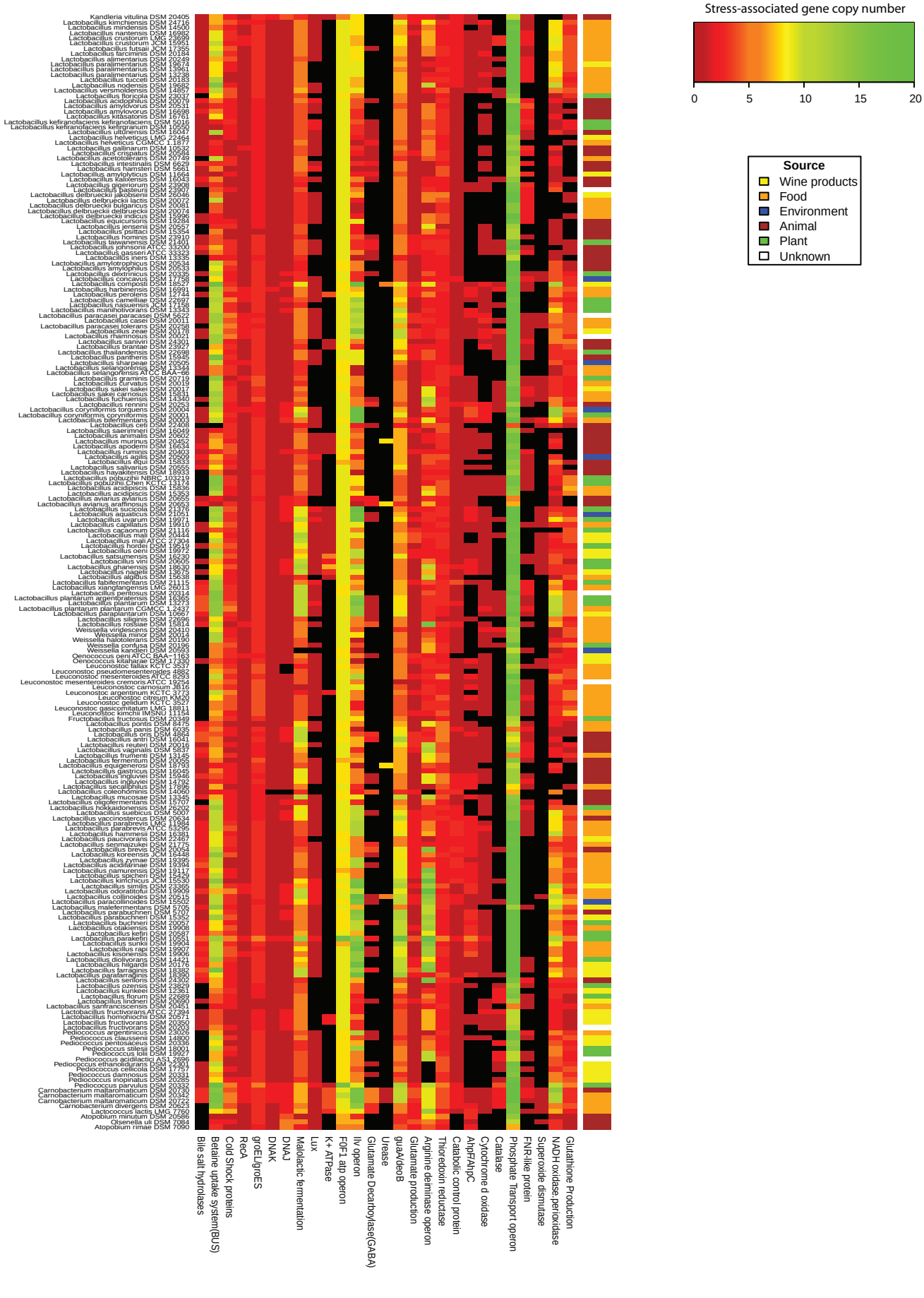




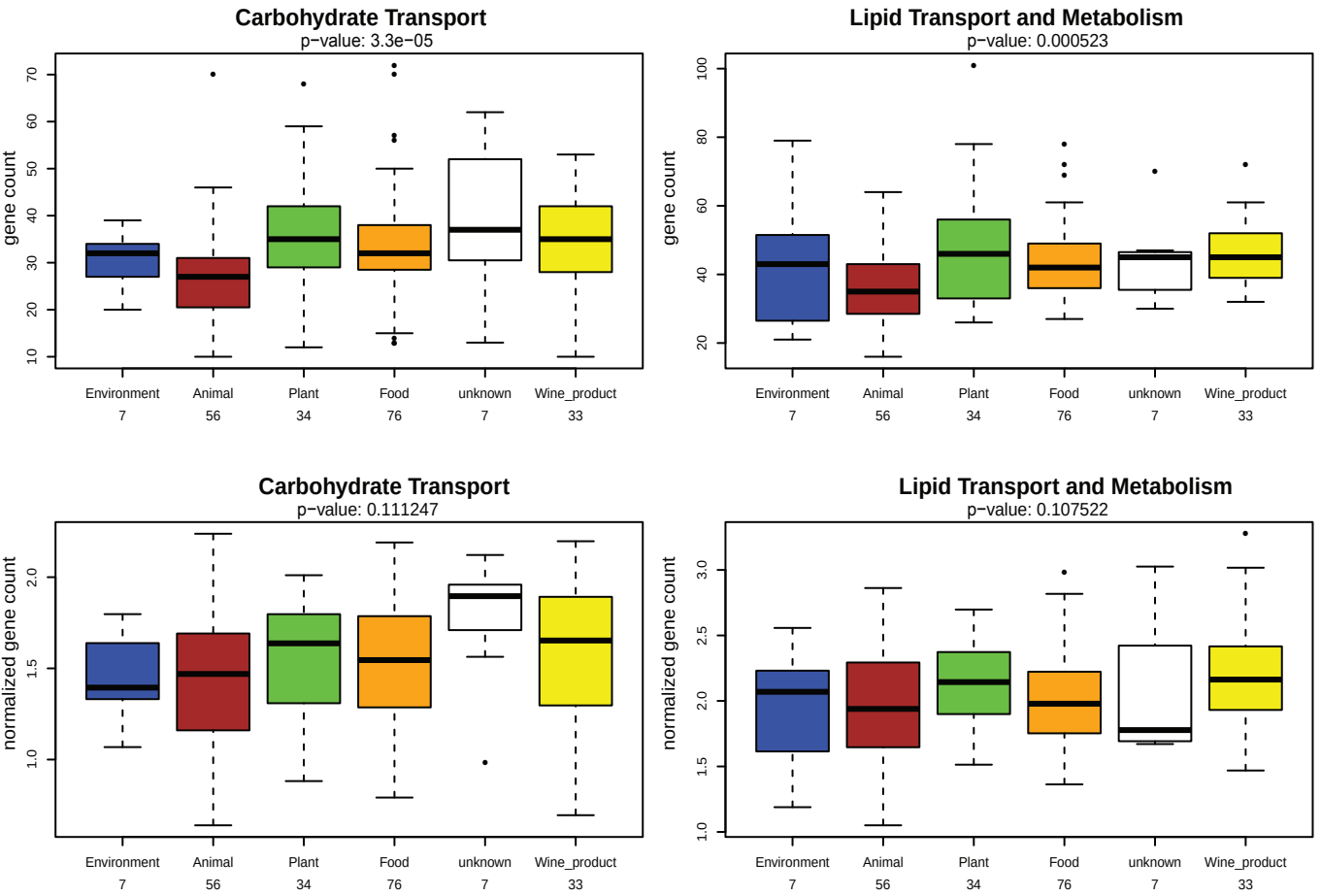




Supp. Figure 23. The effect of normalizing counts of GHs and GTs by genome size. The three bar-graphs on the left show, from top to bottom, GH gene counts, GH gene counts expressed as a percentage of the total gene count and genome size (in kbps). The equivalent for GTs is shown in the three bar-graphs on the right. Genome size is highly correlated with total number of genes per genome (Pearson; 0.99).



Supp. Figure 24. Heatmap of the distribution and abundance of stress response genes across the *Lactobacillus* Genus Complex and associated genera. Gene copy number for 27 stress-associated genes is indicated by the color key from black (absent) to green. Type of stress resistance is indicated by the names at the bottom of the figure. Strains are ordered from top to bottom as they appear top-down on the phylogeny (Fig. 2) with source information for each strain indicated by a color bar at the side of the heat-map.



Supp. Figure 25. Association of carbohydrate transport and lipid transport/metabolism with niche. Top panels display raw gene counts and bottom panels display gene counts normalized by total genes. Boxplots represent a five-point summary of the data in the following order (from bottom to top): minimum, first quartile, median, third quartile and maximum. Outliers are represented as individual points above or below the boxplot.

Supplementary Tables

Supp. Table 1. Species analysed and their genomic features.

Supp. Table 2. Sequence information for the 73 core complete genes. The core genome was established based on genomes listed in Table 1. *Lactobacillus salivarius* genes are shown as exemplars for sequence retrieval. Columns are protein ID numbers, gene designation, locus tag, COG, annotation, co-ordinates in the *L. salivarius* UCC118 genome (NC_007929.1), strand and gene length.

Supp. Table 3. Genera used in building the tree of bacteria.

Supp. Table 4. Sixteen marker genes used to build the bacterial tree composed of 452 genera and 213 *Lactobacillus* genomes and *Lactobacillus* associated genera.

Supp. Table 5. Distribution of LPXTG-containing and sortase enzmyes across the 213 genomes.

Supp. Table 6. **Distribution and abundance of cell envelope proteases (CEPs) and associated anchoring domains and motifs.**

Supp. Table 7. CRISPR occurrence and diversity. CRISPR-Cas system type designation was determined by the presence of CRISPR repeats, spacers, the universal *cas1* gene, and the signature for each type, namely *cas3*, *cas9*, and *cas10*, for Type I, II, and III, respectively. A “Y” in the each of the gene columns designates that the gene was positioned next to the CRISPR locus. In instances where only a partial gene was annotated, the symbol “Y*” is shown. An N indicates that no *cas* gene was found. The type of the CRISPR-Cas system is noted in the “Type” column. When multiple CRISPR loci of the same type were present in the same genome, the type was then designated with a differential letter demarking their order in identification. When *cas* genes could not be annotated, the CRISPR-Cas system type was labeled “undefined.” Strains that contained neither CRISPR repeats nor *cas* genes were labeled with the “N/A” designation. “DR length” corresponds to the number of nucleotides in the CRISPR direct repeat. The number of spacers was also determined for each repeat-spacer array.

Supp. Table 8. Sequence information for the 27 core partial genes.

Supp. Table 9. Presence and absence of the complete pathways for production of the 20 standard amino acids. A = auxotrophic; P = prototrophic.

Supp. Table 10. Presence of sirtuin homologs in the 213 genomes analyzed.

Supp. Table 11. Genomic regions related to bacteriocin production identified in the 213 genomes.

Table S1. Species analyzed and their genome features

Species Name	StrainID	Isolation year	Source	Genome Size (Mbp)	GC%	Predicted ORF number	Sequencing depth	Contig number	Short-read Archive Accession	Genome Accession No.	Niche category	Phylogroup*
Atopobium minutum	DSM-20586	1937	Perineal abscess	1.72	48.69	1595	116.11	64	ERX399734	JQBO00000000	Animal	Other
Atopobium rimae	DSM-7090	1991	Human gingival crevice	1.63	49.26	1524	122.70	9	ERX359753	JQCP00000000	Animal	Other
Carnobacterium divergens	DSM-20623	1984	vacuum packaged minced beef	2.59	34.95	2451	77.13	51	ERX359727	JQBS00000000	Food	Carnobacterium
Carnobacterium maltaromaticum	DSM-20342	1974	Milk with malty flavour	3.76	34.31	3569	53.17	155	ERX359700	JQBG00000000	Food	Carnobacterium
Carnobacterium maltaromaticum	DSM-20722	1974	Vacuum-packaged meat	3.58	34.31	3341	55.86	64	ERX359702	JQBU00000000	Food	Carnobacterium
Carnobacterium maltaromaticum	DSM-20730	1974	diseased rainbow trout	3.54	34.38	3369	56.57	56	ERX399735	JQBV00000000	Animal	Carnobacterium
Fructobacillus fructosus	DSM-20349	1956	Flowers	1.48	44.56	1514	134.69	45	ERX359711	JQBH00000000	Plant	Leu_Fru
Kandleria vitulina	DSM-20405	1973	Calf rumen	2.14	35.03	2126	93.31	138	ERX359782	JQBL00000000	Animal	Other
Lactobacillus acetotolerans	DSM-20749	1986	Fermented Vinegar Broth	1.59	36.26	1531	151.41	123	ERX359757	AYZC00000000	Food	L. delbrueckii
Lactobacillus acidifarinae	DSM-19394	2005	Artisanal wheat sourdough	2.92	51.59	2735	82.17	47	ERX359713	AZDV00000000	Food	L. brevis_L. collinoides
Lactobacillus acidipiscis	DSM-15353	2001	Cheese, Halloumi	2.32	39.06	2280	86.27	298	ERX399738	JQBK00000000	Food	L. salivarius
Lactobacillus acidipiscis	DSM-15836	2000	Fermented fish	2.33	39.07	2307	51.59	457	SRX456282	AZFI00000000	Food	L. salivarius
Lactobacillus acidophilus	DSM-20079	1900	Human	1.95	34.59	1886	61.41	30	SRX456246	AZCS00000000	Animal	L. delbrueckii
Lactobacillus agilis	DSM-20509	1982	Municipal sewage	2.06	41.74	2041	116.27	115	ERX359728	AYYP00000000	Environment	L. salivarius
Lactobacillus algidus	DSM-15638	2000	Vacuum-packed beef	1.59	36.03	1535	150.91	28	SRX456283	AZDI00000000	Food	L. salivarius
Lactobacillus alimentarius	DSM-20249	1970	Marinated fish product	2.34	35.4	2223	102.65	58	ERX359701	AZDQ00000000	Food	L. alimentarius
Lactobacillus amylolyticus	DSM-11664	1999	Acidified beer wort	1.54	38.24	1628	78.05	98	SRX456357	AZEP00000000	Wine product	L. delbrueckii
Lactobacillus amylophilus	DSM-20533	1981	Swine waste-corn fermentation	1.56	43.61	1582	153.98	44	ERX359717	AYYS00000000	Animal	Couple1
Lactobacillus amylotrophicus	DSM-20534	2006	Swine waste-corn fermentation	1.55	43.59	1589	154.98	21	SRX456279	AZCV00000000	Animal	Couple1
Lactobacillus amylovorus	DSM-16698	1981	Faeces, piglet intestine	2.00	37.82	2004	99.92	162	ERX359733	JQBQ00000000	Animal	L. delbrueckii
Lactobacillus amylovorus	DSM-20531	1981	Cattle waste-corn fermentation	2.02	37.77	2061	59.48	116	SRX456280	AZCM00000000	Animal	L. delbrueckii
Lactobacillus animalis	DSM-20602	1983	Dental plaque of baboon	1.89	41.06	1813	127.29	93	ERX359783	AYYW00000000	Animal	L. salivarius
Lactobacillus antri	DSM-16041	2005	Gastric biopsies, Human stomach mucosa	2.24	51.11	2124	53.53	92	SRX456248	AZDK00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus apodemi	DSM-16634	2006	Faeces of wild Japanese wood mouse	2.10	38.63	2009	114.35	109	ERX399737	AZFT00000000	Animal	L. salivarius
Lactobacillus aquaticus	DSM-21051	2009	Surface of a eutrophic freshwater pond	2.41	37.41	2279	99.53	61	ERX359723	AYZD00000000	Environment	L. salivarius
Lactobacillus aviarius subsp. araffinosus	DSM-20653	1986	Intestine of chicken	1.48	38.13	1429	162.40	49	ERX405609	AYYZ00000000	Animal	L. salivarius
Lactobacillus aviarius subsp. aviarius	DSM-20655	1985	Faeces of chicken	1.68	40.12	1591	142.61	41	ERX359760	AYZA00000000	Animal	L. salivarius
Lactobacillus bifermentans	DSM-20003	1943	Blown cheese	3.14	44.29	3088	76.38	205	ERX359761	AZDA00000000	Food	L. coryniformis
Lactobacillus brantae	DSM-23027	2012	Faeces of Canada goose	1.93	47.48	1910	124.36	12	SRX456251	AYZQ00000000	Animal	L. manihotivorans_L. casei
Lactobacillus brevis	DSM-20054	1919	Faeces	2.47	45.96	2425	48.50	11	SRX456278	AZCP00000000	Animal	L. brevis_L. collinoides
Lactobacillus buchneri	DSM-20057	1903	Tomato pulp	2.45	44.41	2349	49.05	90	SRX456296	AZDM00000000	Plant	L. buchneri
Lactobacillus cacaonum	DSM-21116	2009	Cocoa bean heap fermentation	1.92	33.87	1833	124.98	25	ERX399739	AYZE00000000	Plant	L. salivarius
Lactobacillus camelliae	DSM-22697	2007	Fermented tea leaves (miang)	2.57	55.39	2430	93.46	144	ERX359699	AYZJ00000000	Plant	L. manihotivorans_L. casei
Lactobacillus capillatus	DSM-19910	2008	Fermented brine used for stinky tofu production	2.24	37.63	2113	107.36	79	ERX359689	AZEF00000000	Food	L. salivarius
Lactobacillus casei	DSM-20011	1916	Cheese	2.83	46.45	2817	42.46	140	SRX456230	AZCO00000000	Food	L. manihotivorans_L. casei
Lactobacillus ceti	DSM-22408	2008	Lungs of a beaked whale	1.40	33.73	1258	142.46	69	ERX399740	JQBZ00000000	Animal	L. salivarius
Lactobacillus coleohominis	DSM-14060	2001	Human vagina	1.72	40.81	1882	15.69	964	SRX456252	AZEW00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus collinoides	DSM-20515	1972	Fermenting apple juice	3.62	46.11	3380	66.27	165	ERX399741	AYYR00000000	Food	L. brevis_L. collinoides
Lactobacillus composti	DSM-18527	2007	Composting material of distilled shochu residue	3.47	43.95	3370	48.57	131	ERX359694	AZGA00000000	Wine product	L. perolens
Lactobacillus concavus	DSM-17758	2005	Walls of a distilled spirit fermenting cellar	1.90	43.3	1852	85.63	141	ERX399742	AZFX00000000	Environment	Couple2
Lactobacillus coryniformis subsp. coryniformis	DSM-20001	1965	Silage	2.71	42.86	2625	44.36	198	SRX456340	AZCN00000000	Plant	L. coryniformis
Lactobacillus coryniformis subsp. torquens	DSM-20004	1965	Air of cow shed	2.78	42.94	2699	43.11	410	SRX456233	AZDC00000000	Environment	L. coryniformis
Lactobacillus crispatus	DSM-20584	1953	Eye	2.06	36.59	2038	58.34	150	SRX456245	AZCW00000000	Animal	L. delbrueckii
Lactobacillus crustorum	JCM-15951	2007	Wheat sourdough	2.22	35	2150	54.07	101	SRX456344	AZDB00000000	Food	L. alimentarius
Lactobacillus crustorum	LMG-23699	2007	Wheat sourdough	2.24	34.99	2206	89.46	87	ERX359721	JQCK00000000	Food	L. alimentarius
Lactobacillus curvatus	DSM-20019	1903	Milk	1.82	41.97	1819	66.09	250	SRX456225	AZDL00000000	Food	L. sakei
Lactobacillus delbrueckii subsp. bulgaricus	DSM-20081	1919	Bulgarian yoghourt	1.76	49.91	1907	113.73	80	ERX359756	JQAV00000000	Food	L. delbrueckii
Lactobacillus delbrueckii subsp. delbrueckii	DSM-20074	1896	Sour grain mash	1.75	49.86	1884	136.93	69	SRX456341	AZCR00000000	Food	L. delbrueckii
Lactobacillus delbrueckii subsp. indicus	DSM-15996	2005	Traditional dairy fermented product (Dahi type)	1.88	49.54	1871	127.80	239	ERX399745	AZFL00000000	Food	L. delbrueckii
Lactobacillus delbrueckii subsp. jakobsenii	DSM-26046	2013	Dolo wort (Alcholic fermented beverage)	1.75	50.31	1715	102.89	135	SRX690302	JQCG00000000	Wine product	L. delbrueckii
Lactobacillus delbrueckii subsp. lactis	DSM-20072	1919	Emmental cheese	1.87	49.86	1842	64.16	223	SRX456227	AZDE00000000	Food	L. delbrueckii
Lactobacillus dextrinicus	DSM-20335	1964	Silage	1.81	38.05	1743	93.72	36	ERX399746	AYYK00000000	Plant	Couple2
Lactobacillus diolivorans	DSM-14421	2002	Maize silage	3.27	40.01	3119	73.44	154	ERX359710	AZEY00000000	Plant	L. buchneri
Lactobacillus equi	DSM-15833	2002	Faeces of horses	2.30	39.03	2193	104.34	183	ERX359739	AZFH00000000	Animal	L. salivarius
Lactobacillus equicursoris	DSM-19284	2010	Healthy thoroughbred racehorse	2.05	47.71	1887	58.46	316	SRX456237	AZDU00000000	Animal	L. delbrueckii
Lactobacillus equigenerosi	DSM-18793	2008	Faeces of thoroughbred horse	1.60	42.65	1534	149.66	93	ERX359704	AZGC00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus fabifermentans	DSM-21115	2009	Cocoa bean heap fermentation	3.28	45.03	3222	73.09	214	ERX359706	AYGX00000000	Plant	L. plantarum
Lactobacillus farciminis	DSM-20184	1970	Sausage	2.48	36.38	2403	48.37	76	SRX456285	AZDR00000000	Food	L. alimentarius
Lactobacillus farraginis	DSM-18382	2007	Composting material of distilled shochu residue	2.86	42.05	2863	83.78	193	ERX359697	AZFY00000000	Wine product	L. buchneri
Lactobacillus fermentum	DSM-20055	1901	Human saliva	1.90	52.42	1856	105.26	102	ERX359692	JQAU00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus floricola	DSM-23037	2011	Flower of Caltha palustris	1.29	34.53	1240	185.93	30	ERX359719	AYZL00000000	Plant	Single
Lactobacillus florum	DSM-22689	2010	Peony (Paeonia suffruticosa)	1.35	41.14	1317	88.98	50	SRX456305	AYZI00000000	Plant	L. fructivorans
Lactobacillus fructivorans	ATCC-27394	1934	Wine	1.42	38.82	1490	141.19	12	ERX399748	JQAS00000000	Wine product	L. fructivorans
Lactobacillus fructivorans	DSM-20203	1934	N/A	1.37	38.88	1348	87.44	16	SRX456290	AZDS00000000	Unknown	L. fructivorans
Lactobacillus fructivorans	DSM-20350	1934	Spoiled sake	1.38	38.91	1380	144.85	47	ERX399747	JQBI00000000	Wine product	L. fructivorans
Lactobacillus frumenti	DSM-13145	2000	Rye-bran sourdough	1.73	42.55	1681	138.35	34	ERX359759	AZER00000000	Food	L. reuteri_L. vaccinostercus
Lactobacillus fuchuensis	DSM-14340	2002	Vacuum-packaged beef	2.12	41.78	2029	113.00	89	ERX359709	AZEX00000000	Food	L. sakei
Lactobacillus futsaii	JCM-17355	2010	Fu-tsai, a traditional fermented mustard product	2.53	35.6	2482	95.00	208	ERX359725	AZDO00000000	Food	L. alimentarius
Lactobacillus gallinarum	DSM-10532	1992	Chicken crop	1.94	36.48	1956	123.63	113	ERX359738	AZEL00000000	Animal	L. delbrueckii
Lactobacillus gasseri	ATCC-33323	1980	Human	1.89	35.26	1863	NA	1	NA	NC_008530	Animal	L. delbrueckii
Lactobacillus gastricus	DSM-16045	2005	Gastric biopsies, Human stomach mucosa	1.85	41.64	1810	64.92	66	SRX456277	AZFN00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus ghanensis	DSM-18630	2007	Cocoa fermentation	2.61	37.09	2485	92.03	49	ERX359732	AZGB00000000	Plant	L. salivarius
Lactobacillus gigeriorum	DSM-23908	2012	Chicken crop	1.91	36.9	1867	62.96	116	SRX456255	AYZO00000000	Animal	L. delbrueckii
Lactobacillus graminis	DSM-20719	1989	Grass silage	1.84	40.26	1766	130.53	100	ERX359758	AYZB00000000	Plant	L. sakei
Lactobacillus hammesii	DSM-16381	2005	Wheat sourdough	2.82	49.38	2582	85.11	94	ERX359712	AZFS00000000	Food	L. brevis_L. collinoides
Lactobacillus hamsteri	DSM-5661	1988	Faeces of hamster	1.84	35.08	1744	130.37	112	ERX359737	AZGI00000000	Animal	L. delbrueckii
Lactobacillus harbinensis	DSM-16991	2006	Chinese traditional fermented vegetable Suan c	3.14	53.08	3067	76.35	210	ERX399749	AZFW00000000	Food	L. perolens
Lactobacillus hayakitensis	DSM-18933	2007	Faeces of thoroughbred horse	1.70	34.08	1597	141.11	216	ERX359691	AZGD00000000	Animal	L. salivarius
Lactobacillus helveticus	CGMCC-1.1877	1919	Emmental cheese	1.83	36.78	1969	65.62	257	SRX456228	AZEK00000000	Food	L. delbrueckii
Lactobacillus helveticus	LMG-22464	2005	Malt whiskey fermentation	1.77	36.52	1869	112.88	267	ERX359764	JQCJ00000000	Wine product	L. delbrueckii
Lactobacillus hilgardii	DSM-20176	1936	Wine	2.60	39.58	2635	46.10	125	SRX456360	AZDF00000000	Wine product	L. buchneri
Lactobacillus hokkaidonensis	DSM-26202	2013	timothy grass (Phleum pratense L.) silage	2.33	38.13	2285	103.03	84	SRX690303	JQCH00000000	Plant	#L. reuteri_L. vaccinostercus
Lactobacillus hominis	DSM-23910	2013	Human intestine	1.93	35.15	1873	62.29	47	SRX456258	AYZP00000000	Animal	L. delbrueckii
Lactobacillus homohiochii	DSM-20571	1957	Spoiled sake	1.39	38.85	1392	172.28	33	ERX405610	JQBN00000000	Wine product	L. fructivorans
Lactobacillus hordei	DSM-19519	2008	Malted barley	2.30	34.77	2243	104.44	137	ERX359755	AZDX00000000	Plant	L. salivarius
Lactobacillus ineri	DSM-13335	1999	Human urine	1.27	32.52	1186	94.66	58	SRX456259	AZET00000000	Animal	L. delbrueckii
Lactobacillus ingluviei	DSM-14792	2003	Chicken faeces	2.13	49.99	2068	94.05	111	ERX359778	JQBA00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus ingluviei	DSM-15946	2003	Pigeon, crop	2.16	49.88	2101	111.34	113	ERX359729	AZFK00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus intestinalis	DSM-6629	1974	Intestine of rat	2.01	35.37	1858	119.13	87	ERX359740	AZGN00000000	Animal	L. delbrueckii
Lactobacillus jensenii	DSM-20557	1970	Human vaginal discharge	1.61	34.33	1473	74.51	99	SRX456262	AYYU00000000	Animal	L. delbrueckii
Lactobacillus johnsonii	ATCC-33200	1992	Human blood	1.77	34.43	1779	67.66	37	SRX456239	AZCY00000000	Animal	L. delbrueckii
Lactobacillus kalixensis	DSM-16043	2005	Gastric biopsies, Human stomach mucosa	2.08	36.1	1948	115.49	108	ERX359741	AZFM00000000	Animal	L. delbrueckii
Lactobacillus kefiranofaciens subsp. kefiranofa	DSM-5016	1988	Kefir grains	2.26	37.22	2320	53.16	133	SRX456311	AZGG00000000	Plant	L. delbrueckii
Lactobacillus kefiranofaciens subsp. kefirgranu	DSM-10550	1994	Kefir grains	2.10	37.48	2111	114.49	213	ERX359714	AZEM00000000	Plant	L. delbrueckii
Lactobacillus kefiri	DSM-20587	1983	Kefir grains	2.23	41.66	2128	107.57	109	ERX359693	AYYV00000000	Plant	L. buchneri
Lactobacillus kimchicus	JCM-15530	2008	Kimchi	2.59	46.59	2568	92.55	45	SRX456314	AZCX00000000	Food	L. brevis_L. collinoides
Lactobacillus kimchiensis	DSM-24716	2013	Kimchi	2.70	35.48	2577	88.93	71	SRX690301	JQCF00000000	Food	L. alimentarius
Lactobacillus kisonensis	DSM-19906	2009	Sunki, a Japanese traditional pickle	3.01	41.74	2825	39.92	122	SRX456316	AZEB00000000	Food	L. buchneri
Lactobacillus kitasatonis	DSM-16761	2003	Chicken intestine	1.91	37.51	1927	125.47	55	ERX359742	AZFU00000000	Animal	L. delbrueckii
Lactobacillus koreensis	JCM-16448	2011	Cabbage Kimchi	2.97	49.15	2687	80.88	195	ERX359724	AZDP00000000	Food	L. brevis_L. collinoides
Lactobacillus kunkeei	DSM-12361	1998	Commerical grape wine	1.52	36.43	1343	158.28	41	ERX399750	AZCK00000000	Wine product	L. fructivorans
Lactobacillus lindneri	DSM-20690	1901	Spoit beer	1.44	34.14	1437	166.15	49	ERX359707	JQBT00000000	Wine product	L. fructivorans
Lactobacillus malefermentans	DSM-5705	1953	Sour beer	2.05	41.03	2037	58.40	152	SRX456362	AZGJ00000000	Wine product	L. brevis_L. collinoides
Lactobacillus mali	ATCC-27304	1970	Wine must	2.44	36.4	2378	82.07	78	ERX359763	JQAR00000000	Wine product	L. salivarius
Lactobacillus mali	DSM-20444	1970	Apple juice from cider press	2.59	36.06	2538	46.26	211	SRX456318	AYYH00000000	Wine product	L. salivarius
Lactobacillus manihotivorans	DSM-13343	1998	Cassava sour starch fermentation	3.14	47.7	3055	76.42	436	ERX359690	AZEU00000000	Plant	L. manihotivorans_L. casei
Lactobacillus mindensis	DSM-14500	2003	Sourdough	2.34	38.21	2222	102.48	143	ERX359754	AZEZ00000000	Food	L. alimentarius
Lactobacillus mucosae	DSM-13345	2000	Pig small intestine	2.25	46.4	2094	53.23	91	SRX456264	AZEQ00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus murinus	DSM-20452	1982	Intestine of rat	2.20	40.05	2042	83.42	294	ERX359734	AYYN00000000	Animal	L. salivarius
Lactobacillus nagelii	DSM-13675	2000	Partially fermented wine	2.50	36.69	2449	95.93	66	ERX359735	AZEV00000000	Wine product	L. salivarius
Lactobacillus namurensis	DSM-19117	2007	Sourdough	2.48	51.99	2248	96.80	88	ERX399751	AZDT00000000	Food	L. brevis_L. collinoides
Lactobacillus nantensis	DSM-16982	2006	Wheat sourdough	2.91	36.18	2716	82.35	95	SRX456349	AZFV00000000	Food	L. alimentarius
Lactobacillus nasensis	JCM-17158	2012	Sudangrass silage sample	2.28	57.02	2151	105.08	60	ERX359747	AZDJ00000000	Plant	L. manihotivorans_L. casei
Lactobacillus nodensis	DSM-19682	2009	Japanese pickles	2.68	37.57	2639	89.42	28	ERX359744	AZDZ00000000	Food	L. alimentarius
Lactobacillus odoratitofui	DSM-19909	2010	Fermented brine used for stinky tofu production	2.76	44.25	2474	86.99	83	ERX399753	AZEE00000000	Food	L. brevis_L. collinoides
Lactobacillus oeni	DSM-19972	2009	Bobal wine	2.12	37.33	2045	113.19	60	ERX359745	AZEH00000000	Wine product	L. salivarius
Lactobacillus oligofermentans	DSM-15707	2005	Broiler leg	1.83	35.57	1771	130.87	43	ERX359748	AZFE00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus oris	DSM-4864	1988	Human saliva	2.03	49.98	1916	59.06	75	SRX456266	AZGE00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus otakiensis	DSM-19908	2009	Sunki, a Japanese traditional pickle	2.35	42.39	2280	102.25	28	ERX359743	AZED00000000	Food	L. buchneri
Lactobacillus ozensis	DSM-23829	2011	Chrysanthemum, Oze National Park	1.48	31.93	1426	161.73	64	ERX399754	AYYQ00000000	Plant	L. fructivorans
Lactobacillus panis	DSM-6035	1996	Sourdough	2.01	48.05	1915	119.18	198	ERX399755	AZGM00000000	Food	L. reuteri_L. vaccinostercus
Lactobacillus pantheris	DSM-15945	2002	Jaguar faeces	2.55	52.9	2420	94.10	124	ERX359736	AZFJ00000000	Animal	L. manihotivorans_L. casei
Lactobacillus parabrevis	ATCC-53295	2006	Cheese	2.61	49.05	2332	45.90	133	SRX456325	AZCZ00000000	Food	L. brevis_L. collinoides
Lactobacillus parabrevis	LMG-11984	2006	Wheat	2.66	48.98	2398	75.19	123	ERX399756	JQCI00000000	Food	L. brevis_L. collinoides
Lactobacillus parabuchneri	DSM-15352	1989	Blair athol distillery	2.61	43.47	2423	76.63	274	ERX359703	JQBJ00000000	Wine product	L. buchneri
Lactobacillus parabuchneri	DSM-5707	1989	Human saliva	2.57	43.43	2407	93.53	50	SRX456268	AZGK00000000	Animal	L. buchneri
Lactobacillus paracasei subsp. paracasei	DSM-5622	1989	N/A	2.88	46.5	2879	41.64	170	SRX456291	AZGH00000000	Unknown	L. manihotivorans_L. casei
Lactobacillus paracasei subsp. tolerans	DSM-20258	1965	Pasteurized milk	2.38	46.42	2433	50.35	389	SRX456229	AYYJ00000000	Food	L. manihotivorans_L. casei
Lactobacillus paracollinoides	DSM-15502	2004	Brewery environment	3.49	46.84	3247	68.82	275	ERX399758	AZFD00000000	Environment	L. brevis_L. collinoides
Lactobacillus parafarraginis	DSM-18390	2007	Composting material of distilled shochun reside	3.08	45.23	3007	38.94	254	SRX456365	AZFZ00000000	Wine product	L. buchneri
Lactobacillus parakefiri	DSM-10551	1994	Kefir grain	4.91	42.59	4758	40.73	672	ERX399759	AZEN00000000	Plant	L. buchneri
Lactobacillus paralimentarius	DSM-13238	1999	Sourdough	2.55	35.1	2454	94.07	188	ERX359751	AZES00000000	Food	L. alimentarius
Lactobacillus paralimentarius	DSM-13961	1999	Fermented vegetable (Kimchi)	2.75	35.13	2597	87.32	41	ERX399760	AZDH00000000	Food	L. alimentarius
Lactobacillus paralimentarius	DSM-19674	1999	Spanish Bobal grape must	2.89	35.26	2780	83.19	56	ERX359726	AZDY00000000	Wine product	L. alimentarius
Lactobacillus paraplantarum	DSM-10667	1996	Beer contaminant	3.40	43.69	3281	35.34	252	SRX456366	AZEO00000000	Wine product	L. plantarum
Lactobacillus pasteurii	DSM-23907	2013	N/A	1.87	38.51	1794	64.24	44	SRX456269	AYZN00000000	Unknown	L. delbrueckii
Lactobacillus paucivorans	DSM-22467	2010	Yeast storage tank containing lager beer	2.37	49.06	2180	101.18	231	ERX399762	JQCA00000000	Wine product	L. brevis_L. collinoides
Lactobacillus pentosus	DSM-20314	1921	N/A	3.65	46.31	3460	32.90	60	SRX456342	AZCU00000000	Unknown	L. plantarum
Lactobacillus perolens	DSM-12744	2000	Orange lemonade	3.31	49.08	3169	72.53	101	SRX456327	AZEC00000000	Food	L. perolens
Lactobacillus plantarum	DSM-13273	1919	Jojoba meal fermentation	3.45	44.24	3314	58.00	319	ERX399763	JQAW00000000	Plant	L. plantarum
Lactobacillus plantarum subsp. argentoratensis	DSM-16365	2005	Fermented cassava roots (fufu)	3.20	45	3024	74.94	150	ERX359752	AZFR00000000	Plant	L. plantarum
Lactobacillus plantarum subsp. plantarum	CGMCC-1.2437	1919	Pickled cabbage	3.21	44.48	3071	37.40	48	SRX456293	AZEJ00000000	Food	L. plantarum
Lactobacillus pobuzihii	NBRC-103219	2010	Pobuzih (fermented cummingcordia), Cordia di	2.35	37.72	2175	85.27	91	ERX399764	JQCN00000000	Plant	L. salivarius
Lactobacillus pobuzihii.Chen	KCTC-13174	2010	Pobuzih (fermented cummingcordia), Cordia di	2.32	37.71	2144	51.64	103	SRX456329	AZCL00000000	Plant	L. salivarius
Lactobacillus pontis	DSM-8475	1994	Rye sourdough	1.67	53.43	1638	143.66	88	ERX399765	AZGO00000000	Food	L. reuteri_L. vaccinostercus
Lactobacillus psittaci	DSM-15354	2001	Lung of parrott	1.54	35.64	1340	155.70	50	SRX456240	AZFB00000000	Animal	L. delbrueckii
Lactobacillus rapi	DSM-19907	2009	Sunki, a Japanese traditional pickle	2.86	42.95	2729	83.98	109	ERX359746	AZEI00000000	Food	L. buchneri
Lactobacillus rennini	DSM-20253	2006	Rennin	2.27	40.67	2284	105.63	125	ERX399766	AYYI00000000	Animal	L. coryniformis
Lactobacillus reuteri	DSM-20016	1982	Intestine of adult	1.94	38.63	1922	61.98	151	SRX456270	AZDD00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus rhamnosus	DSM-20021	1968	N/A	2.95	46.68	2921	40.69	38	SRX456370	AZCQ00000000	Unknown	L. manihotivorans_L. casei
Lactobacillus rossiae	DSM-15814	2005	Wheat sourdough	2.87	43.34	2743	83.76	61	SRX456353	AZFF00000000	Food	Couple3
Lactobacillus ruminis	DSM-20403	1973	Bovine rumen	2.01	43.39	2061	59.75	115	SRX456242	AYYL00000000	Animal	L. salivarius
Lactobacillus saerimneri	DSM-16049	2004	Pig faeces	1.69	42.57	1669	70.88	92	SRX456271	AZFP00000000	Animal	L. salivarius
Lactobacillus sakei subsp. carnosus	DSM-15831	1996	Fermented meat product	1.99	41.02	1977	120.85	69	ERX399768	AZFG00000000	Food	L. sakei
Lactobacillus sakei subsp. sakei	DSM-20017	1934	"Moto" starter of sake	1.91	41.07	1900	62.82	42	SRX456367	AZDN00000000	Wine product	L. sakei
Lactobacillus salivarius	DSM-20555	1953	Saliva	1.98	32.5	1907	60.52	63	SRX456272	AYYT00000000	Animal	L. salivarius
Lactobacillus sanfranciscensis	DSM-20451	1971	Sourdough	1.23	34.7	1267	97.42	68	SRX456354	AYYM00000000	Food	L. fructivorans
Lactobacillus saniviri	DSM-24301	2012	Faeces of a Japanese healthy adult male	2.44	47.73	2395	98.43	94	ERX399769	JQCE00000000	Animal	L. manihotivorans_L. casei
Lactobacillus satsumensis	DSM-16230	2005	Shochu mash	2.65	39.94	2539	90.73	93	ERX399770	AZFQ00000000	Wine product	L. salivarius
Lactobacillus secaliphilus	DSM-17896	2007	Sourdough	1.65	47.71	1500	121.41	19	ERX359767	JQBW00000000	Food	L. reuteri_L. vaccinostercus
Lactobacillus selangorensis	ATCC-BAA-66	2000	Chili bo	2.18	46.3	2213	91.66	101	ERX359762	JQAT00000000	Food	L. sakei

Lactobacillus selangorensis	DSM-13344	2000	Chili bo	2.09	46.45	2082	95.83	32	SRX690300	JQAZ00000000	Food	L. sakei
Lactobacillus senioris	DSM-24302	2012	Faeces of a healthy 100-year-old Japanese fema	1.57	39.09	1567	153.08	16	ERX359768	AYZR00000000	Animal	L. buchneri
Lactobacillus senmaizukei	DSM-21775	2008	Senmaizuke, a Japanese pickle	2.22	48.64	2149	107.96	68	SRX456331	AYZH00000000	Food	L. brevis_L. collinoides
Lactobacillus sharpeae	DSM-20505	1982	Municipal sewage	2.45	53.38	2371	98.06	79	ERX359770	AYYO00000000	Environment	L. manihotivorans_L. casei
Lactobacillus siliginis	DSM-22696	2006	Wheat sourdough	2.07	44.08	2059	96.60	52	ERX359771	JQCB00000000	Food	Couple3
Lactobacillus similis	DSM-23365	2010	Fermented cane molasses at alcohol plants	3.49	46.99	3206	68.75	282	ERX359772	AYZM00000000	Wine product	L. brevis_L. collinoides
Lactobacillus spicheri	DSM-15429	2004	Rice sourdough	2.75	55.91	2494	87.23	51	ERX359773	AZFC00000000	Food	L. brevis_L. collinoides
Lactobacillus sucicola	DSM-21376	2009	Sap of an Oak tree	2.46	38.48	2335	97.51	31	ERX450947	AYZF00000000	Plant	L. salivarius
Lactobacillus suebicus	DSM-5007	1989	Apple mash	2.65	38.98	2517	45.26	81	SRX456334	AZGF00000000	Food	L. reuteri_L. vaccinostercus
Lactobacillus sunkii	DSM-19904	2009	Sunki, a Japanese traditional pickle	2.69	42.06	2574	89.12	79	SRX456335	AZEA00000000	Food	L. buchneri
Lactobacillus taiwanensis	DSM-21401	2009	Silage cattle feed	1.88	33.97	1851	127.73	93	ERX359776	AYZG00000000	Plant	L. delbrueckii
Lactobacillus thailandensis	DSM-22698	2007	Fermented tea leaves (miang)	2.06	53.5	1943	116.47	23	SRX456337	AYZK00000000	Plant	L. manihotivorans_L. casei
Lactobacillus tucceti	DSM-20183	2009	Sausage	2.17	34.07	2093	110.40	51	ERX359779	AZDG00000000	Food	L. alimentarius
Lactobacillus ultunensis	DSM-16047	2005	Gastric biopsies, Human stomach mucosa	2.16	35.95	2113	55.51	105	SRX456274	AZFO00000000	Animal	L. delbrueckii
Lactobacillus uvarum	DSM-19971	2009	Must of Bobal grape variety	2.69	36.88	2608	89.34	164	ERX359780	AZEG00000000	Plant	L. salivarius
Lactobacillus vaccinostercus	DSM-20634	1983	Cow dung	2.57	43.48	2485	93.56	112	ERX359731	AYYY00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus vaginalis	DSM-5837	1989	Vaginal swab	1.79	40.46	1733	67.13	149	SRX456275	AZGL00000000	Animal	L. reuteri_L. vaccinostercus
Lactobacillus versmoldensis	DSM-14857	2003	Poultry salami	2.37	38.27	2319	50.56	62	SRX456244	AZFA00000000	Food	L. alimentarius
Lactobacillus vini	DSM-20605	2006	Must of grape	2.24	37.54	2191	53.62	269	SRX456339	AYYX00000000	Plant	L. salivarius
Lactobacillus xiangfangensis	LMG-26013	2012	Pickles	3.00	45.1	2806	80.11	145	ERX359765	JQCL00000000	Food	L. plantarum
Lactobacillus zeae	DSM-20178	1959	Corn steep liquor	3.12	47.74	3043	38.45	55	SRX456369	AZCT00000000	Wine product	L. manihotivorans_L. casei
Lactobacillus zymae	DSM-19395	2005	Artisanal wheat sourdough	2.71	53.57	2460	88.66	75	ERX399772	AZDW00000000	Food	#L. brevis_L. collinoides
Lactococcus lactis	LMG-7760	1873	Anchu mash	2.25	35.02	2237	89.07	45	ERX359766	JQCM00000000	Food	Other
Leuconostoc_argentinum	KCTC-3773	1993	Raw milk	1.72	42.89	1821	NA	98	NA	AEGQ00000000	Food	Leu_Fru
Leuconostoc_carnosum	JB16	1989	Kimchi	1.77	37.13	1711	NA	5	NA	CP003851 - CP00385	Food	Leu_Fru
Leuconostoc_citreum	KM20	2008	Kimchi	1.90	38.87	1866	NA	5	NA	DQ489736 - DQ48974	Food	Leu_Fru
Leuconostoc_fallax	KCTC-3537	1992	Sauerkraut	1.64	37.53	1916	NA	30	NA	AEIZ00000000	Food	Leu_Fru
Leuconostoc_gasicomitatum	LMG-18811	2001	Tomato-marinated broiler meat strips	1.95	36.66	1929	NA	1	NA	FN822744	Food	Leu_Fru
Leuconostoc_gelidum	KCTC-3527	1989	Vacuum packaged beef	1.96	36.6	1951	NA	43	NA	AEMI00000000	Food	Leu_Fru
Leuconostoc_kimchii	IMSNU-11154	2000	Kimchi	2.10	37.91	2110	NA	6	NA	CP001753 - CP00175	Food	Leu_Fru
Leuconostoc_mesenteroides	ATCC-8293	1878	Fermenting olives	2.08	37.67	2056	NA	2	NA	C_008496, NC_0085	Food	Leu_Fru
Leuconostoc_mesenteroides_cremoris	ATCC-19254	1929	Hansen's dried starter powder	1.74	37.9	1791	NA	29	NA	CZKK01	Unknown	Leu_Fru
Leuconostoc_pseudomesenteroides	4882	N/A	N/A	2.01	39.06	2180	NA	106	NA	CAKV00000000	Food	Leu_Fru
Oenococcus_kitaharae	DSM-17330	2006	Distilled residue of shochu mashes	1.84	42.68	1900	NA	2	NA	AFVZ00000000	Wine product	Oenococcus
Oenococcus_oeni	ATCC-BAA-1163	2002	Fermented beverages	1.75	37.94	2055	NA	62	NA	AAUV00000000	Wine product	Oenococcus
Olsenella uli	DSM-7084	1991	Human gingival crevice	2.06	64.69	1848	97.22	13	ERX399771	JQCO00000000	Animal	Other
Pediococcus acidilactici	AS1.2696	N/A	N/A	1.93	42.13	1849	51.85	18	SRX689743	JQAQ00000000	Unknown	Pediococcus
Pediococcus argentinicus	DSM-23026	2008	Fermented wheat flour	1.76	36.67	1772	56.66	93	SRX689746	JQCQ00000000	Food	Pediococcus
Pediococcus cellicola	DSM-17757	2005	Distilled piri-fermenting cellar	2.04	39.04	1974	49.06	22	SRX689747	JQBR00000000	Wine product	Pediococcus
Pediococcus claussenii	DSM-14800	2002	Spolied beer	1.88	36.74	1807	53.21	44	SRX689748	JQBB00000000	Wine product	Pediococcus
Pediococcus damnosus	DSM-20331	1903	Lager beer yeast	2.19	38.23	2085	45.63	201	SRX689749	JQBD00000000	Wine product	Pediococcus
Pediococcus ethanolidurans	DSM-22301	2006	Walls of a distilled-spirit-fermenting cellar	2.26	37.18	2180	44.23	66	SRX689750	JQBY00000000	Wine product	Pediococcus
Pediococcus inopinatus	DSM-20285	1988	Brewery yeast	2.11	38.61	2081	47.30	157	SRX689751	JQBC00000000	Wine product	Pediococcus
Pediococcus lolii	DSM-19927	1887	Ryegrass silage	2.04	42.13	1971	49.00	31	SRX689752	JQCC00000000	Plant	Pediococcus
Pediococcus parvulus	DSM-20332	1961	Silage	3.99	40.38	3917	25.04	153	SRX689754	JQBE00000000	Plant	Pediococcus
Pediococcus pentosaceus	DSM-20336	1934	Dried American beer yeast	1.74	37.25	1687	57.49	28	SRX689755	JQBF00000000	Wine product	Pediococcus
Pediococcus stilesii	DSM-18001	2006	White maize grains	1.84	38.11	1834	54.41	47	SRX689756	JQBX00000000	Plant	Pediococcus
Weissella confusa	DSM-20196	1969	Sugar cane	2.22	44.73	2075	90.16	40	ERX359705	JQAY00000000	Plant	Weissella
Weissella halotolerans	DSM-20190	1983	Sausage	1.37	43.06	1341	146.41	12	ERX359698	JQAX00000000	Food	Weissella
Weissella kandleri	DSM-20593	1983	Desert spring	1.33	39.67	1281	149.88	21	ERX359708	JQBP00000000	Environment	Weissella
Weissella minor	DSM-20014	1983	Milking machine slime	1.77	39.29	1777	112.85	59	ERX399773	JQCD00000000	Food	Weissella
Weissella viridescens	DSM-20410	1957	Cured meat products	1.54	41.09	1525	130.11	8	ERX359781	JQBM00000000	Food	Weissella

Table S2. Sequence information for the 73 core genes

PID*	Gene	locus tag*	COG	Annotation	Co-ordinates*	Strand*	Length*
90960992	*dnaN*	LSL_0002	COG0592L	DNA polymerase III subunit beta	1532..2671	+	1140
90960995	*gyrB*	LSL_0005	COG0187L	DNA gyrase subunit B	4451..6409	+	1959
90960996	*gyrA*	LSL_0006	COG0188L	DNA gyrase subunit A	6446..8998	+	2553
90960997	*rpsF*	LSL_0007	COG0360J	30S ribosomal protein S6	9217..9507	+	291
90960998	*ssb*	LSL_0008	COG0629L	Single-strand DNA binding protein	9548..10099	+	552
90960999	*rpsR*	LSL_0009	COG0238J	30S ribosomal protein S18	10121..10357	+	237
90961180	*rpoC*	LSL_0198	COG0086K	DNA-directed RNA polymerase subunit beta'	238653..242318	+	3666
90961182	*rpsL*	LSL_0200	COG0048J	30S ribosomal protein S12	243364..243777	+	414
90961184	*efg*	LSL_0202	COG0480J	elongation factor G	244403..246496	+	2094
90961200	*trpS*	LSL_0218	COG0180J	Tryptophanyl-tRNA synthetase II	271973..272992	+	1020
90961335	*murF*	LSL_0355	COG0770M	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	387857..389227	+	1371
90961453	-	LSL_0477	COG0537FGR	bis(5'-nucleosyl)-tetraphosphatase	525874..526302	-	429
90961460	-	LSL_0484	COG0073R	tRNA-binding domain-containing protein	530707..531357	+	651
90961464	*polA*	LSL_0488	COG0749L	DNA polymerase I	536545..539223	+	2679
90961470	*thrS*	LSL_0494	COG0441J	threonyl-tRNA synthetase	543759..545708	+	1950
90961471	*infC*	LSL_0495	COG0290J	translation initiation factor IF-3	545916..546440	+	525
90961480	-	LSL_0504	COG0799S	Iojap-related protein	552036..552389	+	354
90961487	*rpsB*	LSL_0511	COG0052J	30S ribosomal protein S2	557234..558031	+	798
90961488	*tsf*	LSL_0512	COG0264J	elongation factor Ts	558125..559000	+	876
90961519	-	LSL_0543	COG0218R	GTP-binding protein	584473..585063	+	591
90961537	*pyrH*	LSL_0561	COG0528F	uridylate kinase	600230..600952	+	723
90961538	*frr*	LSL_0562	COG0233J	ribosome recycling factor	600955..601518	+	564
90961539	*uppS*	LSL_0563	COG0020I	undecaprenyl pyrophosphate synthetase	601653..602435	+	783
90961540	*cdsA*	LSL_0564	COG0575I	phosphatidate cytidylyltransferase	602438..603226	+	789
90961544	-	LSL_0568	COG0779S	hypothetical protein	610934..611407	+	474
90961545	*nusA*	LSL_0569	COG0195K	transcription elongation factor NusA	611433..612560	+	1128
90961567	-	LSL_0591	COG2890J	peptide release factor-glutamine N5-methyltransferase	634747..635586	+	840
90961569	*upp*	LSL_0593	COG0035F	uracil phosphoribosyltransferase	636708..637337	+	630
90961575	*atpG*	LSL_0599	COG0224C	F0F1 ATP synthase subunit gamma	641185..642114	+	930
90961576	*atpD*	LSL_0600	COG0055C	F0F1 ATP synthase subunit beta	642139..643545	+	1407
90961630	*typA*	LSL_0653	COG1217T	GTP-binding protein	697034..698875	+	1842
90961790	*pheS*	LSL_0813	COG0016J	phenylalanyl-tRNA synthetase subunit alpha	829184..830230	+	1047
90961836	*pfs*	LSL_0859	COG0775F	5'-methylthioadenosine nucleosidase	878731..879417	+	687
90962022	*ftsZ*	LSL_1047	COG0206D	cell division protein FtsZ	1070613..1071866	-	1254
90962026	*murD*	LSL_1051	COG0771M	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase	1075412..1076788	-	1377
90962027	*mraY*	LSL_1052	COG0472M	phospho-N-acetylmuramoyl-pentapeptide-transferase	1076817..1077785	-	969
90962042	*mreC*	LSL_1067	COG1792M	rod shape-determining protein MreC	1089421..1090275	-	855
90962072	*obgE*	LSL_1097	COG0536R	GTPase ObgE	1125579..1126877	-	1299
90962084	-	LSL_1109	COG0816L	Holliday junction resolvase-like protein	1136102..1136533	-	432
90962100	*ruvB*	LSL_1125	COG2255L	Holliday junction DNA helicase RuvB	1153506..1154516	-	1011
90962101	*ruvA*	LSL_1126	COG0632L	Holliday junction DNA helicase RuvA	1154555..1155160	-	606
90962106	*pgsA*	LSL_1131	COG0558I	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	1163042..1163629	-	588
90962110	*ftsK*	LSL_1135	COG1674D	cell division protein	1168139..1170418	-	2280
90962125	-	LSL_1150	COG0802R	ATP/GTP hydrolase	1183489..1183941	-	453
90962130	*smpB*	LSL_1155	COG0691O	SsrA-binding protein	1187273..1187740	-	468
90962160	*prfB*	LSL_1185	COG1186J	peptide chain release factor 2	1222388..1223455	-	1068
90962186	*groEL*	LSL_1211	COG0459O	molecular chaperone GroEL	1246385..1248007	-	1623
90962187	*groS*	LSL_1212	COG0234O	molecular chaperone GroES	1248037..1248321	-	285
90962191	*gcp*	LSL_1216	COG0533O	O-sialoglycoprotein endopeptidase	1252076..1253107	-	1032
90962198	*holB*	LSL_1223	COG2812L	DNA polymerase III subunit delta'	1257561..1258553	-	993
90962199	*tmk*	LSL_1224	COG0125F	thymidylate kinase	1258588..1259214	-	627
90962204	-	LSL_1229	COG0590FJ	cytosine/adenosine deaminase	1262491..1262991	-	501
90962212	*rplJ*	LSL_1238	COG0244J	50S ribosomal protein L10	1271311..1271814	-	504
90962215	*nusG*	LSL_1241	COG0250K	transcription antitermination protein	1273402..1273962	-	561
90962219	*spoU*	LSL_1245	COG0566J	tRNA/rRNA methyltransferase	1275013..1275759	-	747
90962221	*cysS*	LSL_1247	COG0215J	cysteinyl-tRNA synthetase	1276159..1277571	-	1413
90962278	*ppnK*	LSL_1304	COG0061G	inorganic polyphosphate/ATP-NAD kinase	1351254..1352060	-	807
90962327	-	LSL_1355	COG0037D	tRNA(Ile)-lysidine synthase TilS	1418710..1420068	-	1359
90962332	*mfd*	LSL_1360	COG1197LK	transcription-repair coupling factor	1422969..1426493	-	3525
90962333	*pth*	LSL_1361	COG0193J	peptidyl-tRNA hydrolase	1426515..1427072	-	558
90962374	*rplM*	LSL_1403	COG0102J	50S ribosomal protein L13	1478312..1478755	-	444
90962376	*cbiQ*	LSL_1405	COG0619P	cobalt transport permease	1479676..1480470	-	795
90962380	*rpoA*	LSL_1409	COG0202K	DNA-directed RNA polymerase subunit alpha	1482888..1483832	-	945
90962387	*rplO*	LSL_1416	COG0200J	50S ribosomal protein L15	1487364..1487798	-	435
90962389	*rpsE*	LSL_1418	COG0098J	30S ribosomal protein S5	1488030..1488530	-	501
90962391	*rplF*	LSL_1420	COG0097J	50S ribosomal protein L6	1488953..1489489	-	537
90962392	*rpsH*	LSL_1421	COG0096J	30S ribosomal protein S8	1489522..1489920	-	399
90962394	*rplE*	LSL_1423	COG0094J	50S ribosomal protein L5	1490153..1490695	-	543
90962400	*rpsC*	LSL_1429	COG0092J	30S ribosomal protein S3	1492402..1493058	-	657
90962404	*rplW*	LSL_1433	COG0089J	50S ribosomal protein L23	1494625..1494909	-	285
90962405	*rplD*	LSL_1434	COG0088J	50S ribosomal protein L4	1494909..1495532	-	624
90962406	*rplC*	LSL_1435	COG0087J	50S ribosomal protein L3	1495557..1496180	-	624
90962696	*rplI*	LSL_1727	COG0359J	50S ribosomal protein L9	1809124..1809573	-	450

*These columns are provided according to the reference genome*L. salivarius* UCC118

Table S3. Genera used in building the tree of bacteria

Accession No.	Phylum	Class	Order	Family	Genus	Species
NC_012483	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	*Acidobacterium*	*Acidobacterium capsulatum*
NC_014963	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	*Terriglobus*	*Terriglobus saanensis*
NC_015064	Acidobacteria	Acidobacteria	Acidobacteriales	Acidobacteriaceae	*Granulicella*	*Granulicella tundricola*
NC_008536	Acidobacteria	Solibacteres	Solibacterales	Solibacteraceae	*Candidatus Solibacter*	*Candidatus Solibacter usitatus*
NC_008009	Acidobacteria	Unclassified	Unclassified	Unclassified	*Candidatus Koribacter*	*Candidatus Koribacter versatilis*
NC_000962	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	*Mycobacterium*	*Mycobacterium tuberculosis*
NC_002935	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	*Corynebacterium*	*Corynebacterium diphtheriae*
NC_003155	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	*Streptomyces*	*Streptomyces avermitilis*
NC_004307	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	*Bifidobacterium*	*Bifidobacterium longum*
NC_004551	Actinobacteria	Actinobacteria	Actinomycetales	Unclassified	*Tropheryma*	*Tropheryma whipplei*
NC_006085	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	*Propionibacterium*	*Propionibacterium acnes*
NC_006087	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	*Leifsonia*	*Leifsonia xyli*
NC_006361	Actinobacteria	Actinobacteria	Actinomycetales	Nocardiaceae	*Nocardia*	*Nocardia farcinica*
NC_007333	Actinobacteria	Actinobacteria	Actinomycetales	Nocardiopsaceae	*Thermobifida*	*Thermobifida fusca*
NC_007777	Actinobacteria	Actinobacteria	Actinomycetales	Frankiaceae	*Frankia*	*Unclassified*
NC_008148	Actinobacteria	Actinobacteria	Rubrobacterales	Rubrobacteraceae	*Rubrobacter*	*Rubrobacter xylanophilus*
NC_008268	Actinobacteria	Actinobacteria	Actinomycetales	Nocardiaceae	*Rhodococcus*	*Rhodococcus jostii*
NC_008541	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Arthrobacter*	*Unclassified*
NC_008578	Actinobacteria	Actinobacteria	Actinomycetales	Acidothermaceae	*Acidothermus*	*Acidothermus cellulolyticus*
NC_008699	Actinobacteria	Actinobacteria	Actinomycetales	Nocardioidaceae	*Nocardioides*	*Unclassified*
NC_009142	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Saccharopolyspora*	*Saccharopolyspora erythraea*
NC_009380	Actinobacteria	Actinobacteria	Actinomycetales	Micromonosporaceae	*Salinispora*	*Salinispora tropica*
NC_009480	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	*Clavibacter*	*Clavibacter michiganensis*
NC_009664	Actinobacteria	Actinobacteria	Actinomycetales	Kineosporiaceae	*Kineococcus*	*Kineococcus radiotolerans*
NC_010168	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Renibacterium*	*Renibacterium salmoninarum*
NC_010617	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Kocuria*	*Kocuria rhizophila*
NC_012669	Actinobacteria	Actinobacteria	Actinomycetales	Beutenbergiaceae	*Beutenbergia*	*Beutenbergia cavernae*
NC_012803	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Micrococcus*	*Micrococcus luteus*
NC_013093	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Actinosynnema*	*Actinosynnema mirum*
NC_013124	Actinobacteria	Actinobacteria	Acidimicrobiales	Acidimicrobiaceae	*Acidimicrobium*	*Acidimicrobium ferrooxidans*
NC_013131	Actinobacteria	Actinobacteria	Actinomycetales	Catenulisporaceae	*Catenulispora*	*Catenulispora acidiphila*
NC_013159	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Saccharomonospora*	*Saccharomonospora viridis*
NC_013165	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Slackia*	*Slackia heliotrinireducens*
NC_013169	Actinobacteria	Actinobacteria	Actinomycetales	Dermacoccaceae	*Kytococcus*	*Kytococcus sedentarius*
NC_013170	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Cryptobacterium*	*Cryptobacterium curtum*
NC_013172	Actinobacteria	Actinobacteria	Actinomycetales	Dermabacteraceae	*Brachybacterium*	*Brachybacterium faecium*
NC_013174	Actinobacteria	Actinobacteria	Actinomycetales	Jonesiaceae	*Jonesia*	*Jonesia denitrificans*
NC_013203	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Atopobium*	*Atopobium parvulum*
NC_013204	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Eggerthella*	*Eggerthella lenta*
NC_013235	Actinobacteria	Actinobacteria	Actinomycetales	Nakamurellaceae	*Nakamurella*	*Nakamurella multipartita*
NC_013510	Actinobacteria	Actinobacteria	Actinomycetales	Thermomonosporaceae	*Thermomonospora*	*Thermomonospora curvata*
NC_013521	Actinobacteria	Actinobacteria	Actinomycetales	Sanguibacteraceae	*Sanguibacter*	*Sanguibacter keddieii*
NC_013530	Actinobacteria	Actinobacteria	Actinomycetales	Promicromonosporaceae	*Xylanimonas*	*Xylanimonas cellulosilytica*
NC_013595	Actinobacteria	Actinobacteria	Actinomycetales	Streptosporangiaceae	*Streptosporangium*	*Streptosporangium roseum*
NC_013715	Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	*Rothia*	*Rothia mucilaginosa*
NC_013721	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	*Gardnerella*	*Gardnerella vaginalis*
NC_013739	Actinobacteria	Actinobacteria	Solirubrobacterales	Conexibacteraceae	*Conexibacter*	*Conexibacter woesei*
NC_013947	Actinobacteria	Actinobacteria	Actinomycetales	Glycomycetaceae	*Stackebrandtia*	*Stackebrandtia nassauensis*
NC_014151	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	*Cellulomonas*	*Cellulomonas flavigena*
NC_014158	Actinobacteria	Actinobacteria	Actinomycetales	Tsukamurellaceae	*Tsukamurella*	*Tsukamurella paurometabola*
NC_014165	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Thermobispora*	*Thermobispora bispora*
NC_014168	Actinobacteria	Actinobacteria	Actinomycetales	Segniliparaceae	*Segniliparus*	*Segniliparus rotundus*
NC_014218	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Arcanobacterium*	*Arcanobacterium haemolyticum*
NC_014246	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Mobiluncus*	*Mobiluncus curtisii*
NC_014318	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Amycolatopsis*	*Amycolatopsis mediterranei*
NC_014363	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Olsenella*	*Olsenella uli*
NC_014830	Actinobacteria	Actinobacteria	Actinomycetales	Intrasporangiaceae	*Intrasporangium*	*Intrasporangium calvum*
NC_015125	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	*Microbacterium*	*Microbacterium testaceum*
NC_015312	Actinobacteria	Actinobacteria	Actinomycetales	Pseudonocardiaceae	*Pseudonocardia*	*Pseudonocardia dioxanivorans*
NC_015389	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	*Coriobacterium*	*Coriobacterium glomerans*
NC_015564	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	*Amycolicicoccus*	*Amycolicicoccus subflavus*
NC_015588	Actinobacteria	Actinobacteria	Actinomycetales	Promicromonosporaceae	*Isoptericola*	*Isoptericola variabilis*
NC_015635	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	*Microlunatus*	*Microlunatus phosphovorus*
NC_000918	Aquificae	Aquificae	Aquificales	Aquificaceae	*Aquifex*	*Aquifex aeolicus*
NC_010730	Aquificae	Aquificae	Aquificales	Hydrogenothermaceae	*Sulfurihydrogenibium*	*Unclassified*
NC_011126	Aquificae	Aquificae	Aquificales	Aquificaceae	*Hydrogenobaculum*	*Unclassified*
NC_012440	Aquificae	Aquificae	Aquificales	Hydrogenothermaceae	*Persephonella*	*Persephonella marina*
NC_013799	Aquificae	Aquificae	Aquificales	Aquificaceae	*Hydrogenobacter*	*Hydrogenobacter thermophilus*
NC_013894	Aquificae	Aquificae	Aquificales	Aquificaceae	*Thermocrinis*	*Thermocrinis albus*
NC_014926	Aquificae	Aquificae	Aquificales	Desulfurobacteriaceae	*Thermovibrio*	*Thermovibrio ammonificans*
NC_015185	Aquificae	Aquificae	Aquificales	Desulfurobacteriaceae	*Desulfurobacterium*	*Desulfurobacterium thermolithotrophum*
NC_002950	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	*Porphyromonas*	*Porphyromonas gingivalis*
NC_003228	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	*Bacteroides*	*Bacteroides fragilis*
NC_009615	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	*Parabacteroides*	*Parabacteroides distasonis*
NC_011565	Bacteroidetes	Bacteroidia	Bacteroidales	Unclassified	*Candidatus Azobacteroides*	*Candidatus Azobacteroides pseudotrichonymphae*
NC_014033	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	*Prevotella*	*Prevotella ruminicola*
NC_014734	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	*Paludibacter*	*Paludibacter propionicigenes*
NC_015160	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	*Odoribacter*	*Odoribacter splanchnicus*
NC_008255	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Cytophaga*	*Cytophaga hutchinsonii*
NC_013037	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Dyadobacter*	*Dyadobacter fermentans*
NC_013730	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Spirosoma*	*Spirosoma linguale*
NC_014655	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Leadbetterella*	*Leadbetterella byssophila*
NC_014759	Bacteroidetes	Cytophagia	Cytophagales	Flammeovirgaceae	*Marivirga*	*Marivirga tractuosa*
NC_015703	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Runella*	*Runella slithyformis*
NC_015914	Bacteroidetes	Cytophagia	Cytophagales	Cyclobacteriaceae	*Cyclobacterium*	*Cyclobacterium marinum*
NC_008571	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Gramella*	*Gramella forsetii*
NC_009441	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Flavobacterium*	*Flavobacterium johnsoniae*

NC_013123	Bacteroidetes	Flavobacteria	Flavobacteriales	Unclassified	*Candidatus Sulcia*	*Candidatus Sulcia muelleri*
NC_013162	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Capnocytophaga*	*Capnocytophaga ochracea*
NC_013222	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Robiginitalea*	*Robiginitalea biformata*
NC_013418	Bacteroidetes	Flavobacteria	Flavobacteriales	Blattabacteriaceae	*Blattabacterium*	*Blattabacterium sp. (Periplaneta americana)*
NC_014041	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Zunongwangia*	*Zunongwangia profunda*
NC_014230	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Croceibacter*	*Croceibacter atlanticus*
NC_014472	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Maribacter*	*Unclassified*
NC_014738	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Riemerella*	*Riemerella anatipestifer*
NC_014934	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Cellulophaga*	*Cellulophaga algicola*
NC_015144	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Weeksella*	*Weeksella virosa*
NC_015321	Bacteroidetes	Flavobacteria	Flavobacteriales	Cryomorphaceae	*Fluviicola*	*Fluviicola taffensis*
NC_015496	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Krokinobacter*	*Unclassified*
NC_015638	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Lacinutrix*	*Unclassified*
NC_015844	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Zobellia*	*Unclassified*
NC_015945	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	*Muricauda*	*Muricauda ruestringensis*
NC_013061	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Sphingobacteriaceae	*Pedobacter*	*Pedobacter heparinus*
NC_013132	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Unclassified	*Chitinophaga*	*Chitinophaga pinensis*
NC_015277	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Sphingobacteriaceae	*Sphingobacterium*	*Unclassified*
NC_015510	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Saprospiraceae	*Haliscomenobacter*	*Haliscomenobacter hydrossis*
NC_007677	Bacteroidetes	Unclassified	Bacteroidetes Order II.	Rhodothermaceae	*Salinibacter*	*Salinibacter ruber*
NC_010830	Bacteroidetes	Unclassified	Unclassified	Unclassified	*Candidatus Amoebophilus*	*Candidatus Amoebophilus asiaticus*
NC_015966	Bacteroidetes	Unclassified	Bacteroidetes Order II.	Rhodothermaceae	*Rhodothermus*	*Rhodothermus marinus*
NC_000117	Chlamydiae	Chlamydiia	Chlamydiales	Chlamydiaceae	*Chlamydia*	*Chlamydia trachomatis*
NC_003361	Chlamydiae	Chlamydiia	Chlamydiales	Chlamydiaceae	*Chlamydophila*	*Chlamydophila caviae*
NC_005861	Chlamydiae	Chlamydiia	Chlamydiales	Parachlamydiaceae	*Candidatus Protochlamydia*	*Candidatus Protochlamydia amoebophila*
NC_014225	Chlamydiae	Chlamydiia	Chlamydiales	Waddliaceae	*Waddlia*	*Waddlia chondrophila*
NC_015702	Chlamydiae	Chlamydiia	Chlamydiales	Parachlamydiaceae	*Parachlamydia*	*Parachlamydia acanthamoebae*
NC_015713	Chlamydiae	Chlamydiia	Chlamydiales	Simkaniaceae	*Simkania*	*Simkania negevensis*
NC_002932	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Chlorobaculum*	*Chlorobaculum tepidum*
NC_007514	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Chlorobium*	*Chlorobium chlorochromatii*
NC_011026	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Chloroherpeton*	*Chloroherpeton thalassium*
NC_011059	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Prosthecochloris*	*Prosthecochloris aestuarii*
NC_011060	Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	*Pelodictyon*	*Pelodictyon phaeoclathratiforme*
NC_014960	Chloroflexi	Anaerolineae	Anaerolineales	Anaerolineaceae	*Anaerolinea*	*Anaerolinea thermophila*
NC_009523	Chloroflexi	Chloroflexi	Chloroflexales	Chloroflexaceae	*Roseiflexus*	*Unclassified*
NC_009972	Chloroflexi	Chloroflexi	Herpetosiphonales	Herpetosiphonaceae	*Herpetosiphon*	*Herpetosiphon aurantiacus*
NC_010175	Chloroflexi	Chloroflexi	Chloroflexales	Chloroflexaceae	*Chloroflexus*	*Chloroflexus aurantiacus*
NC_002936	Chloroflexi	Dehalococcoidia	Dehalococcoidales	Dehalococcoidaceae	*Dehalococcoides*	*Dehalococcoides mccartyi*
NC_014314	Chloroflexi	Dehalococcoidia	Unclassified	Unclassified	*Dehalogenimonas*	*Dehalogenimonas lykanthroporepellens*
NC_011959	Chloroflexi	Thermomicrobia	Thermomicrobiales	Thermomicrobiaceae	*Thermomicrobium*	*Thermomicrobium roseum*
NC_013523	Chloroflexi	Thermomicrobia	Sphaerobacterales	Sphaerobacteraceae	*Sphaerobacter*	*Sphaerobacter thermophilus*
NC_014836	Chrysiogenetes	Chrysiogenetes	Chrysiogenales	Chrysiogenaceae	*Desulfurispirillum*	*Desulfurispirillum indicum*
NC_005125	Cyanobacteria	Gloeobacteria	Gloeobacterales	Unclassified	*Gloeobacter*	*Gloeobacter violaceus*
NC_003272	Cyanobacteria	Unclassified	Nostocales	Nostocaceae	*Nostoc*	*Unclassified*
NC_004113	Cyanobacteria	Unclassified	Chroococcales	Unclassified	*Thermosynechococcus*	*Thermosynechococcus elongatus*
NC_005042	Cyanobacteria	Unclassified	Prochlorales	Prochlorococcaceae	*Prochlorococcus*	*Prochlorococcus marinus*
NC_005070	Cyanobacteria	Unclassified	Chroococcales	Unclassified	*Synechococcus*	*Unclassified*
NC_007413	Cyanobacteria	Unclassified	Nostocales	Nostocaceae	*Anabaena*	*Anabaena variabilis*
NC_008312	Cyanobacteria	Unclassified	Oscillatoriales	Unclassified	*Trichodesmium*	*Trichodesmium erythraeum*
NC_011726	Cyanobacteria	Unclassified	Chroococcales	Unclassified	*Cyanothece*	*Unclassified*
NC_014248	Cyanobacteria	Unclassified	Nostocales	Nostocaceae	*Trichormus*	*Trichormus azollae*
NC_013939	Deferribacteres	Deferribacteres	Deferribacterales	Deferribacteraceae	*Deferribacter*	*Deferribacter desulfuricans*
NC_013943	Deferribacteres	Deferribacteres	Deferribacterales	Deferribacteraceae	*Denitrovibrio*	*Denitrovibrio acetiphilus*
NC_014758	Deferribacteres	Deferribacteres	Deferribacterales	Deferribacteraceae	*Calditerrivibrio*	*Calditerrivibrio nitroreducens*
NC_015672	Deferribacteres	Deferribacteres	Deferribacterales	Deferribacteraceae	*Flexistipes*	*Flexistipes sinusarabici*
NC_001263	Deinococcus-Therm	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	*Deinococcus radiodurans*
NC_005835	Deinococcus-Therm	Deinococci	Thermales	Thermaceae	*Thermus*	*Thermus thermophilus*
NC_013946	Deinococcus-Therm	Deinococci	Thermales	Thermaceae	*Meiothermus*	*Meiothermus ruber*
NC_014221	Deinococcus-Therm	Deinococci	Deinococcales	Trueperaceae	*Truepera*	*Truepera radiovictrix*
NC_014761	Deinococcus-Therm	Deinococci	Thermales	Thermaceae	*Oceanithermus*	*Oceanithermus profundus*
NC_015387	Deinococcus-Therm	Deinococci	Thermales	Thermaceae	*Marinithermus*	*Marinithermus hydrothermalis*
NC_011297	Dictyoglomi	Dictyoglomia	Dictyoglomales	Dictyoglomaceae	*Dictyoglomus*	*Dictyoglomus thermophilum*
NC_010644	Elusimicrobia	Elusimicrobia	Elusimicrobiales	Elusimicrobiaceae	*Elusimicrobium*	*Elusimicrobium minutum*
NC_013410	Fibrobacteres	Fibrobacteria	Fibrobacterales	Fibrobacteraceae	*Fibrobacter*	*Fibrobacter succinogenes*
NC_000964	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	*Bacillus subtilis*
NC_002662	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Lactococcus*	*Lactococcus lactis*
NC_002737	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	*Streptococcus*	*Streptococcus pyogenes*
NC_002745	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	*Staphylococcus*	*Staphylococcus aureus*
NC_004193	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Oceanobacillus*	*Oceanobacillus iheyensis*
NC_004567	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Lactobacillus*	*Lactobacillus plantarum*
NC_006510	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Geobacillus*	*Geobacillus kaustophilus*
NC_008525	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	*Pediococcus*	*Pediococcus pentosaceus*
NC_008528	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	*Oenococcus*	*Oenococcus oeni*
NC_008531	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	*Leuconostoc*	*Leuconostoc mesenteroides*
NC_008555	Firmicutes	Bacilli	Bacillales	Listeriaceae	*Listeria*	*Listeria welshimeri*
NC_010556	Firmicutes	Bacilli	Bacillales	Unclassified	*Exiguobacterium*	*Exiguobacterium sibiricum*
NC_011567	Firmicutes	Bacilli	Bacillales	Bacillaceae	*Anoxybacillus*	*Anoxybacillus flavithermus*
NC_011999	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	*Macrococcus*	*Macrococcus caseolyticus*
NC_012491	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	*Brevibacillus*	*Brevibacillus brevis*
NC_012914	Firmicutes	Bacilli	Bacillales	Paenibacillaceae	*Paenibacillus*	*Unclassified*
NC_013205	Firmicutes	Bacilli	Bacillales	Alicyclobacillaceae	*Alicyclobacillus*	*Alicyclobacillus acidocaldarius*
NC_014098	Firmicutes	Bacilli	Bacillales	Alicyclobacillaceae	*Kyrpidia*	*Kyrpidia tusciae*
NC_015278	Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	*Aerococcus*	*Aerococcus urinae*
NC_015391	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	*Carnobacterium*	*Unclassified*
NC_015516	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	*Melissococcus*	*Melissococcus plutonius*
NC_015759	Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	*Weissella*	*Weissella koreensis*
NC_003030	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*Clostridium*	*Clostridium acetobutylicum*
NC_003869	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Caldanaerobacter*	*Caldanaerobacter subterraneus*
NC_006177	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVIII. In	*Symbiobacterium*	*Symbiobacterium thermophilum*
NC_007503	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Carboxydothermus*	*Carboxydothermus hydrogenoformans*

NC_007644	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Moorella*	*Moorella thermoacetica*
NC_007907	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Desulfitobacterium*	*Desulfitobacterium hafniense*
NC_008346	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	*Syntrophomonas*	*Syntrophomonas wolfei*
NC_009253	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Desulfotomaculum*	*Desulfotomaculum reducens*
NC_009437	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacterales Fami	*Caldicellulosiruptor*	*Caldicellulosiruptor saccharolyticus*
NC_009454	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Pelotomaculum*	*Pelotomaculum thermopropionicum*
NC_009633	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*Alkaliphilus*	*Alkaliphilus metalliredigens*
NC_010320	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Thermoanaerobacter*	*Unclassified*
NC_010337	Firmicutes	Clostridia	Clostridiales	Heliobacteriaceae	*Heliobacterium*	*Heliobacterium modesticaldum*
NC_010376	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XI. Incer	*Finegoldia*	*Finegoldia magna*
NC_010424	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Candidatus Desulforudis*	*Candidatus Desulforudis audaxviator*
NC_010718	Firmicutes	Clostridia	Natranaerobiales	Natranaerobiaceae	*Natranaerobius*	*Natranaerobius thermophilus*
NC_011295	Firmicutes	Clostridia	Thermoanaerobacterale	Thermodesulfobiaceae	*Coprothermobacter*	*Coprothermobacter proteolyticus*
NC_011899	Firmicutes	Clostridia	Halanaerobiales	Halanaerobiaceae	*Halothermothrix*	*Halothermothrix orenii*
NC_012781	Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	*Eubacterium*	*Eubacterium rectale*
NC_013171	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XI. Incer	*Anaerococcus*	*Anaerococcus prevotii*
NC_013385	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Ammonifex*	*Ammonifex degensii*
NC_014152	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Thermincola*	*Thermincola potens*
NC_014220	Firmicutes	Clostridia	Clostridiales	Syntrophomonadaceae	*Syntrophothermus*	*Syntrophothermus lipocalidus*
NC_014377	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacterales Fami	*Thermosediminibacter*	*Thermosediminibacter oceani*
NC_014378	Firmicutes	Clostridia	Halanaerobiales	Halobacteroidaceae	*Acetohalobium*	*Acetohalobium arabaticum*
NC_014387	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Butyrivibrio*	*Butyrivibrio proteoclasticus*
NC_014410	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacterales Fami	*Thermoanaerobacterium*	*Thermoanaerobacterium thermosaccharolyticum*
NC_014654	Firmicutes	Clostridia	Halanaerobiales	Halanaerobiaceae	*Halanaerobium*	*Unclassified*
NC_014828	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	*Ethanoligenens*	*Ethanoligenens harbinense*
NC_014831	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVII. Inc	*Thermaerobacter*	*Thermaerobacter marianensis*
NC_014833	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	*Ruminococcus*	*Ruminococcus albus*
NC_015172	Firmicutes	Clostridia	Clostridiales	Peptococcaceae	*Syntrophobotulus*	*Syntrophobotulus glycolicus*
NC_015275	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Cellulosilyticum*	*Cellulosilyticum lentocellum*
NC_015499	Firmicutes	Clostridia	Thermoanaerobacterale	Thermodesulfobiaceae	*Thermodesulfobium*	*Thermodesulfobium narugense*
NC_015519	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacteraceae	*Tepidanaerobacter*	*Tepidanaerobacter acetatoxydans*
NC_015520	Firmicutes	Clostridia	Thermoanaerobacterale	Thermoanaerobacterales Fami	*Mahella*	*Mahella australiensis*
NC_015757	Firmicutes	Clostridia	Clostridiales	Clostridiales Family XVII. Inc	*Sulfobacillus*	*Sulfobacillus acidophilus*
NC_015913	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	*Candidatus Arthromitus*	*Candidatus Arthromitus sp. SFB-mouse*
NC_015977	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Roseburia*	*Roseburia hominis*
NC_015601	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	*Erysipelothrix*	*Erysipelothrix rhusiopathiae*
NC_013520	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	*Veillonella*	*Veillonella parvula*
NC_013740	Firmicutes	Negativicutes	Selenomonadales	Acidaminococcaceae	*Acidaminococcus*	*Acidaminococcus fermentans*
NC_015437	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	*Selenomonas*	*Selenomonas sputigena*
NC_003454	Fusobacteria	Fusobacteria	Fusobacteriales	Fusobacteriaceae	*Fusobacterium*	*Fusobacterium nucleatum*
NC_013192	Fusobacteria	Fusobacteria	Fusobacteriales	Leptotrichiaceae	*Leptotrichia*	*Leptotrichia buccalis*
NC_013515	Fusobacteria	Fusobacteria	Fusobacteriales	Leptotrichiaceae	*Streptobacillus*	*Streptobacillus moniliformis*
NC_014632	Fusobacteria	Fusobacteria	Fusobacteriales	Fusobacteriaceae	*Ilyobacter*	*Ilyobacter polytropus*
NC_012489	Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadaceae	*Gemmatimonas*	*Gemmatimonas aurantiaca*
NC_011296	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	*Thermodesulfovibrio*	*Thermodesulfovibrio yellowstonii*
NC_014355	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	*Nitrospira*	*Unclassified*
NC_005027	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	*Rhodopirellula*	*Rhodopirellula baltica*
NC_013720	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	*Pirellula*	*Pirellula staleyi*
NC_014148	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	*Planctomyces*	*Planctomyces limnophilus*
NC_014962	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	*Isosphaera*	*Isosphaera pallida*
NC_000963	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	*Rickettsia*	*Rickettsia prowazekii*
NC_002678	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Mesorhizobium*	*Mesorhizobium loti*
NC_002696	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	*Caulobacter*	*Caulobacter vibrioides*
NC_003047	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Sinorhizobium*	*Sinorhizobium meliloti*
NC_003317	Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	*Brucella*	*Brucella melitensis*
NC_003911	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Ruegeria*	*Ruegeria pomeroyi*
NC_004463	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	*Bradyrhizobium*	*Bradyrhizobium diazoefficiens*
NC_005295	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	*Ehrlichia*	*Ehrlichia ruminantium*
NC_005296	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	*Rhodopseudomonas*	*Rhodopseudomonas palustris*
NC_005955	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	*Bartonella*	*Bartonella quintana*
NC_006526	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Zymomonas*	*Zymomonas mobilis*
NC_006677	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Gluconobacter*	*Gluconobacter oxydans*
NC_007205	Proteobacteria	Alphaproteobacteria	Unclassified	Unclassified	*Candidatus Pelagibacter*	*Candidatus Pelagibacter ubique*
NC_007406	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	*Nitrobacter*	*Nitrobacter winogradskyi*
NC_007493	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Rhodobacter*	*Rhodobacter sphaeroides*
NC_007626	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	*Magnetospirillum*	*Magnetospirillum magneticum*
NC_007643	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	*Rhodospirillum*	*Rhodospirillum rubrum*
NC_007722	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	*Erythrobacter*	*Erythrobacter litoralis*
NC_007761	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Rhizobium*	*Rhizobium etli*
NC_007794	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Novosphingobium*	*Novosphingobium aromaticivorans*
NC_007797	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	*Anaplasma*	*Anaplasma phagocytophilum*
NC_007798	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	*Neorickettsia*	*Neorickettsia sennetsu*
NC_007802	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Jannaschia*	*Unclassified*
NC_008048	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Sphingopyxis*	*Sphingopyxis alaskensis*
NC_008209	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Roseobacter*	*Roseobacter denitrificans*
NC_008254	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Chelativorans*	*Unclassified*
NC_008343	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Granulibacter*	*Granulibacter bethesdensis*
NC_008347	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Hyphomonadaceae	*Maricaulis*	*Maricaulis maris*
NC_008358	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Hyphomonadaceae	*Hyphomonas*	*Hyphomonas neptunium*
NC_008576	Proteobacteria	Alphaproteobacteria	Magnetococcales	Magnetococcaceae	*Magnetococcus*	*Magnetococcus marinus*
NC_009484	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Acidiphilium*	*Acidiphilium cryptum*
NC_009488	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	*Orientia*	*Orientia tsutsugamushi*
NC_009511	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Sphingomonas*	*Sphingomonas wittichii*
NC_009667	Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	*Ochrobactrum*	*Ochrobactrum anthropi*
NC_009719	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhodobiaceae	*Parvibaculum*	*Parvibaculum lavamentivorans*
NC_009720	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	*Xanthobacter*	*Xanthobacter autotrophicus*
NC_009937	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	*Azorhizobium*	*Azorhizobium caulinodans*
NC_009952	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Dinoroseobacter*	*Dinoroseobacter shibae*
NC_010125	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Gluconacetobacter*	*Gluconacetobacter diazotrophicus*
NC_010172	Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	*Methylobacterium*	*Methylobacterium extorquens*

NC_010581	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	*Beijerinckia*	*Beijerinckia indica*
NC_011144	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	*Phenylobacterium*	*Phenylobacterium zucineum*
NC_011386	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	*Oligotropha*	*Oligotropha carboxidovorans*
NC_011666	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	*Methylocella*	*Methylocella silvestris*
NC_011985	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Agrobacterium*	*Agrobacterium tumefaciens*
NC_012982	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Hyphomonadaceae	*Hirschia*	*Hirschia baltica*
NC_012985	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	*Candidatus Liberibacter*	*Candidatus Liberibacter asiaticus*
NC_013209	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	*Acetobacter*	*Acetobacter pasteurianus*
NC_014006	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	*Sphingobium*	*Sphingobium japonicum*
NC_014010	Proteobacteria	Alphaproteobacteria	Unclassified	Unclassified	*Candidatus Puniceispirillum*	*Candidatus Puniceispirillum marinum*
NC_014217	Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	*Starkeya*	*Starkeya novella*
NC_014313	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	*Hyphomicrobium*	*Hyphomicrobium denitrificans*
NC_014375	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	*Brevundimonas*	*Brevundimonas subvibrioides*
NC_014414	Proteobacteria	Alphaproteobacteria	Parvularculales	Parvularculaceae	*Parvularcula*	*Parvularcula bermudensis*
NC_014664	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	*Rhodomicrobium*	*Rhodomicrobium vannielii*
NC_002927	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Bordetella*	*Bordetella bronchiseptica*
NC_002946	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Neisseria*	*Neisseria gonorrhoeae*
NC_003295	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Ralstonia*	*Ralstonia solanacearum*
NC_004757	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	*Nitrosomonas*	*Nitrosomonas europaea*
NC_005085	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Chromobacterium*	*Chromobacterium violaceum*
NC_006350	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Burkholderia*	*Burkholderia pseudomallei*
NC_006513	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*Aromatoleum*	*Aromatoleum aromaticum*
NC_007298	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*Dechloromonas*	*Dechloromonas aromatica*
NC_007404	Proteobacteria	Betaproteobacteria	Hydrogenophilales	Hydrogenophilaceae	*Thiobacillus*	*Thiobacillus denitrificans*
NC_007614	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	*Nitrosospira*	*Nitrosospira multiformis*
NC_007908	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Albidiferax*	*Albidiferax ferrireducens*
NC_007947	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	*Methylobacillus*	*Methylobacillus flagellatus*
NC_007948	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Polaromonas*	*Unclassified*
NC_007973	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Cupriavidus*	*Cupriavidus metallidurans*
NC_008702	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*Azoarcus*	*Unclassified*
NC_008752	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Acidovorax*	*Acidovorax citrulli*
NC_008786	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Verminephrobacter*	*Verminephrobacter eiseniae*
NC_008825	Proteobacteria	Betaproteobacteria	Burkholderiales	Unclassified	*Methylibium*	*Methylibium petroleiphilum*
NC_009379	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	*Polynucleobacter*	*Polynucleobacter necessarius*
NC_010002	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Delftia*	*Delftia acidovorans*
NC_010524	Proteobacteria	Betaproteobacteria	Burkholderiales	Unclassified	*Leptothrix*	*Leptothrix cholodnii*
NC_011662	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	*Thauera*	*Unclassified*
NC_012559	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Laribacter*	*Laribacter hongkongensis*
NC_012791	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Variovorax*	*Variovorax paradoxus*
NC_012968	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	*Methylotenera*	*Methylotenera mobilis*
NC_012969	Proteobacteria	Betaproteobacteria	Methylophilales	Methylophilaceae	*Methylovorus*	*Methylovorus glucosotrophus*
NC_013194	Proteobacteria	Betaproteobacteria	Unclassified	Unclassified	*Candidatus Accumulibacter*	*Candidatus Accumulibacter phosphatis*
NC_013446	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Comamonas*	*Comamonas testosteroni*
NC_013959	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	*Sideroxydans*	*Sideroxydans lithotrophicus*
NC_014153	Proteobacteria	Betaproteobacteria	Burkholderiales	Unclassified	*Thiomonas*	*Thiomonas intermedia*
NC_014323	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Herbaspirillum*	*Herbaspirillum seropedicae*
NC_014394	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	*Gallionella*	*Gallionella capsiferriformans*
NC_014640	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Achromobacter*	*Achromobacter xylosoxidans*
NC_014910	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Alicycliphilus*	*Alicycliphilus denitrificans*
NC_014914	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Taylorella*	*Taylorella equigenitalis*
NC_015458	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Pusillimonas*	*Unclassified*
NC_015677	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Ramlibacter*	*Ramlibacter tataouinensis*
NC_015856	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Collimonas*	*Collimonas fungivorans*
NC_016002	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	*Pseudogulbenkiania*	*Unclassified*
NC_002937	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	*Desulfovibrio*	*Desulfovibrio vulgaris*
NC_002939	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Geobacteraceae	*Geobacter*	*Geobacter sulfurreducens*
NC_005363	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionaceae	*Bdellovibrio*	*Bdellovibrio bacteriovorus*
NC_007498	Proteobacteria	Deltaproteobacteria	Desulfuromonadales	Pelobacteraceae	*Pelobacter*	*Pelobacter carbinolicus*
NC_007759	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophaceae	*Syntrophus*	*Syntrophus aciditrophicus*
NC_007760	Proteobacteria	Deltaproteobacteria	Myxococcales	Myxococcaceae	*Anaeromyxobacter*	*Anaeromyxobacter dehalogenans*
NC_008011	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	*Lawsonia*	*Lawsonia intracellularis*
NC_008095	Proteobacteria	Deltaproteobacteria	Myxococcales	Myxococcaceae	*Myxococcus*	*Myxococcus xanthus*
NC_008554	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophobacteraceae	*Syntrophobacter*	*Syntrophobacter fumaroxidans*
NC_009943	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobacteraceae	*Desulfococcus*	*Desulfococcus oleovorans*
NC_010162	Proteobacteria	Deltaproteobacteria	Myxococcales	Polyangiaceae	*Sorangium*	*Sorangium cellulosum*
NC_011768	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobacteraceae	*Desulfatibacillum*	*Desulfatibacillum alkenivorans*
NC_012108	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobacteraceae	*Desulfobacterium*	*Desulfobacterium autotrophicum*
NC_013173	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfomicrobiaceae	*Desulfomicrobium*	*Desulfomicrobium baculatum*
NC_013223	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfohalobiaceae	*Desulfohalobium*	*Desulfohalobium retbaense*
NC_013440	Proteobacteria	Deltaproteobacteria	Myxococcales	Kofleriaceae	*Haliangium*	*Haliangium ochraceum*
NC_014216	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	*Desulfurivibrio*	*Desulfurivibrio alkaliphilus*
NC_014365	Proteobacteria	Deltaproteobacteria	Desulfarculales	Desulfarculaceae	*Desulfarculus*	*Desulfarculus baarsii*
NC_014972	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	*Desulfobulbus*	*Desulfobulbus propionicus*
NC_015318	Proteobacteria	Deltaproteobacteria	Desulfurellales	Desulfurellaceae	*Hippea*	*Hippea maritima*
NC_015388	Proteobacteria	Deltaproteobacteria	Syntrophobacterales	Syntrophaceae	*Desulfobacca*	*Desulfobacca acetoxidans*
NC_002163	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	*Campylobacter*	*Campylobacter jejuni*
NC_007575	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	*Sulfurimonas*	*Sulfurimonas denitrificans*
NC_009662	Proteobacteria	Epsilonproteobacteria	Unclassified	Unclassified	*Nitratiruptor*	*Unclassified*
NC_009663	Proteobacteria	Epsilonproteobacteria	Unclassified	Unclassified	*Sulfurovum*	*Unclassified*
NC_009850	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	*Arcobacter*	*Arcobacter butzleri*
NC_011333	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	*Helicobacter*	*Helicobacter pylori*
NC_012115	Proteobacteria	Epsilonproteobacteria	Nautiliales	Nautiliaceae	*Nautilia*	*Nautilia profundicola*
NC_013512	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	*Sulfurospirillum*	*Sulfurospirillum deleyianum*
NC_014762	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	*Sulfuricurvum*	*Sulfuricurvum kujiense*
NC_014935	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Unclassified	*Nitratifractor*	*Nitratifractor salsuginis*
NC_000907	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Haemophilus*	*Haemophilus influenzae*
NC_000913	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Escherichia*	*Escherichia coli*
NC_002488	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Xylella*	*Xylella fastidiosa*
NC_002516	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Pseudomonas*	*Pseudomonas aeruginosa*
NC_002528	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Buchnera*	*Buchnera aphidicola*

NC_002663	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Pasteurella*	*Pasteurella multocida*
NC_002971	Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	*Coxiella*	*Coxiella burnetii*
NC_003143	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Yersinia*	*Yersinia pestis*
NC_003197	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Salmonella*	*Salmonella enterica*
NC_003902	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Xanthomonas*	*Xanthomonas campestris*
NC_003910	Proteobacteria	Gammaproteobacteria	Alteromonadales	Colwelliaceae	*Colwellia*	*Colwellia psychrerythraea*
NC_004337	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Shigella*	*Shigella flexneri*
NC_004344	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Wigglesworthia*	*Wigglesworthia glossinidia*
NC_004347	Proteobacteria	Gammaproteobacteria	Alteromonadales	Shewanellaceae	*Shewanella*	*Shewanella oneidensis*
NC_004459	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	*Vibrio*	*Vibrio vulnificus*
NC_004547	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Pectobacterium*	*Pectobacterium atrosepticum*
NC_005061	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Candidatus Blochmannia*	*Unclassified*
NC_005126	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Photorhabdus*	*Photorhabdus luminescens*
NC_005966	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	*Unclassified*
NC_006300	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Basfia*	*Mannheimia succiniciproducens*
NC_006368	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae	*Legionella*	*Legionella pneumophila*
NC_006370	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	*Photobacterium*	*Photobacterium profundum*
NC_006512	Proteobacteria	Gammaproteobacteria	Alteromonadales	Idiomarinaceae	*Idiomarina*	*Idiomarina loihiensis*
NC_006570	Proteobacteria	Gammaproteobacteria	Thiotrichales	Francisellaceae	*Francisella*	*Francisella tularensis*
NC_006840	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	*Aliivibrio*	*Aliivibrio fischeri*
NC_007204	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Psychrobacter*	*Psychrobacter arcticus*
NC_007481	Proteobacteria	Gammaproteobacteria	Alteromonadales	Pseudoalteromonadaceae	*Pseudoalteromonas*	*Pseudoalteromonas haloplanktis*
NC_007484	Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	*Nitrosococcus*	*Nitrosococcus oceani*
NC_007520	Proteobacteria	Gammaproteobacteria	Thiotrichales	Piscirickettsiaceae	*Thiomicrospira*	*Thiomicrospira crunogena*
NC_007645	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Hahellaceae	*Hahella*	*Hahella chejuensis*
NC_007912	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	*Saccharophagus*	*Saccharophagus degradans*
NC_007963	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	*Chromohalobacter*	*Chromohalobacter salexigens*
NC_008260	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	*Alcanivorax*	*Alcanivorax borkumensis*
NC_008309	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Histophilus*	*Histophilus somni*
NC_008340	Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	*Alkalilimnicola*	*Alkalilimnicola ehrlichii*
NC_008570	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	*Aeromonas*	*Aeromonas hydrophila*
NC_008709	Proteobacteria	Gammaproteobacteria	Alteromonadales	Psychromonadaceae	*Psychromonas*	*Psychromonas ingrahamii*
NC_008740	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	*Marinobacter*	*Marinobacter hydrocarbonoclasticus*
NC_008789	Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	*Halorhodospira*	*Halorhodospira halophila*
NC_009053	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Actinobacillus*	*Actinobacillus pleuropneumoniae*
NC_009436	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Enterobacter*	*Unclassified*
NC_009648	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Klebsiella*	*Klebsiella pneumoniae*
NC_009654	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	*Marinomonas*	*Unclassified*
NC_009778	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Cronobacter*	*Cronobacter sakazakii*
NC_009792	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Citrobacter*	*Citrobacter koseri*
NC_009832	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Serratia*	*Serratia proteamaculans*
NC_010554	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Proteus*	*Proteus mirabilis*
NC_010694	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Erwinia*	*Erwinia tasmaniensis*
NC_010943	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Stenotrophomonas*	*Stenotrophomonas maltophilia*
NC_010995	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Cellvibrio*	*Cellvibrio japonicus*
NC_011206	Proteobacteria	Gammaproteobacteria	Acidithiobacillales	Acidithiobacillaceae	*Acidithiobacillus*	*Acidithiobacillus ferrooxidans*
NC_012560	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Azotobacter*	*Azotobacter vinelandii*
NC_012691	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	*Tolumonas*	*Tolumonas auensis*
NC_012751	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Candidatus Hamiltonella*	*Candidatus Hamiltonella defensa*
NC_012880	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Dickeya*	*Dickeya dadantii*
NC_012913	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Aggregatibacter*	*Aggregatibacter aphrophilus*
NC_012997	Proteobacteria	Gammaproteobacteria	Alteromonadales	Unclassified	*Teredinibacter*	*Teredinibacter turnerae*
NC_013166	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	*Kangiella*	*Kangiella koreensis*
NC_013422	Proteobacteria	Gammaproteobacteria	Chromatiales	Halothiobacillaceae	*Halothiobacillus*	*Halothiobacillus neapolitanus*
NC_013851	Proteobacteria	Gammaproteobacteria	Chromatiales	Chromatiaceae	*Allochromatium*	*Allochromatium vinosum*
NC_013889	Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	*Thioalkalivibrio*	*Unclassified*
NC_013892	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Xenorhabdus*	*Xenorhabdus bovienii*
NC_013956	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Pantoea*	*Pantoea ananatis*
NC_014109	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Candidatus Riesia*	*Candidatus Riesia pediculicola*
NC_014147	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Moraxella*	*Moraxella catarrhalis*
NC_014532	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	*Halomonas*	*Halomonas elongata*
NC_014541	Proteobacteria	Gammaproteobacteria	Alteromonadales	Ferrimonadaceae	*Ferrimonas*	*Ferrimonas balearica*
NC_014924	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	*Pseudoxanthomonas*	*Pseudoxanthomonas suwonensis*
NC_015061	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Rahnella*	*Unclassified*
NC_015460	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	*Gallibacterium*	*Gallibacterium anatis*
NC_015497	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	*Glaciecola*	*Unclassified*
NC_015554	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	*Alteromonas*	*Unclassified*
NC_015572	Proteobacteria	Gammaproteobacteria	Methylococcales	Methylococcaceae	*Methylomonas*	*Methylomonas methanica*
NC_015581	Proteobacteria	Gammaproteobacteria	Thiotrichales	Piscirickettsiaceae	*Thioalkalimicrobium*	*Thioalkalimicrobium cyclicum*
NC_015735	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Candidatus Moranella*	*Candidatus Moranella endobia*
NC_000919	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Treponema*	*Treponema pallidum*
NC_001318	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Borrelia*	*Borrelia burgdorferi*
NC_004342	Spirochaetes	Spirochaetia	Spirochaetales	Leptospiraceae	*Leptospira*	*Leptospira interrogans*
NC_012225	Spirochaetes	Spirochaetia	Spirochaetales	Brachyspiraceae	*Brachyspira*	*Brachyspira hyodysenteriae*
NC_014364	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Spirochaeta*	*Spirochaeta smaragdinae*
NC_015152	Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	*Sphaerochaeta*	*Sphaerochaeta globosa*
NC_013522	Synergistetes	Synergistia	Synergistales	Synergistaceae	*Thermanaerovibrio*	*Thermanaerovibrio acidaminovorans*
NC_014011	Synergistetes	Synergistia	Synergistales	Synergistaceae	*Aminobacterium*	*Aminobacterium colombiense*
NC_000908	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	*Mycoplasma*	*Mycoplasma genitalium*
NC_002162	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	*Ureaplasma*	*Ureaplasma parvum*
NC_005303	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	*Candidatus Phytoplasma*	*Onion yellows phytoplasma*
NC_006055	Tenericutes	Mollicutes	Entomoplasmatales	Entomoplasmataceae	*Mesoplasma*	*Mesoplasma florum*
NC_010163	Tenericutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	*Acholeplasma*	*Acholeplasma laidlawii*
NC_015681	Thermodesulfobacte	Thermodesulfobacteria	Thermodesulfobacterial	Thermodesulfobacteriaceae	*Thermodesulfatator*	*Thermodesulfatator indicus*
NC_015682	Thermodesulfobacte	Thermodesulfobacteria	Thermodesulfobacterial	Thermodesulfobacteriaceae	*Thermodesulfobacterium*	*Thermodesulfobacterium geofontis*
NC_000853	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Thermotoga*	*Thermotoga maritima*
NC_009616	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Thermosipho*	*Thermosipho melanesiensis*
NC_009718	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Fervidobacterium*	*Fervidobacterium nodosum*
NC_010003	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Petrotoga*	*Petrotoga mobilis*
NC_012785	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	*Kosmotoga*	*Kosmotoga olearia*

NC_013525	Unclassified	Unclassified	Unclassified	Unclassified	*Thermobaculum*	*Thermobaculum terrenum*
NC_010571	Verrucomicrobia	Opitutae	Unclassified	Opitutaceae	*Opitutus*	*Opitutus terrae*
NC_014008	Verrucomicrobia	Opitutae	Puniceicoccales	Puniceicoccaceae	*Coraliomargarita*	*Coraliomargarita akajimensis*
NC_010794	Verrucomicrobia	Unclassified	Methylacidiphilales	Methylacidiphilaceae	*Methylacidiphilum*	*Methylacidiphilum infernorum*
NC_010655	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	*Akkermansia*	*Akkermansia muciniphila*

Table S4. List of genes used in building the tree of bacteria

PID	Gene	Locus tag*	COG	Predicted product	Position*	Length (bp)*
90961179	*rpoB*	LSL_0197	COG0085K	DNA-directed RNA polymerase subunit beta	235025..238624	3600
90961471	*infC*	LSL_0495	COG0290J	translation initiation factor IF-3	545916..546440	525
90961487	*rpsB*	LSL_0511	COG0052J	30S ribosomal protein S2	557234..558031	798
90961488	*tsf*	LSL_0512	COG0264J	elongation factor Ts	558125..559000	876
90961538	*frr*	LSL_0562	COG0233J	ribosome recycling factor	600955..601518	564
90961545	*nusA*	LSL_0569	COG0195K	transcription elongation factor NusA	611433..612560	1128
90962130	*smpB*	LSL_1155	COG0691O	SsrA-binding protein	1187273..1187740	468
90962374	*rplM*	LSL_1403	COG0102J	50S ribosomal protein L13	1478312..1478755	444
90962389	*rpsE*	LSL_1418	COG0098J	30S ribosomal protein S5	1488030..1488530	501
90962391	*rplF*	LSL_1420	COG0097J	50S ribosomal protein L6	1488953..1489489	537
90962394	*rplE*	LSL_1423	COG0094J	50S ribosomal protein L5	1490153..1490695	543
90962399	*rplP*	LSL_1428	COG0197J	50S ribosomal protein L16	1491964..1492398	435
90962400	*rpsC*	LSL_1429	COG0092J	30S ribosomal protein S3	1492402..1493058	657
90962403	*rplB*	LSL_1432	COG0090J	50S ribosomal protein L2	1493766..1494599	834
90962405	*rplD*	LSL_1434	COG0088J	50S ribosomal protein L4	1494909..1495532	624
90962406	*rplC*	LSL_1435	COG0087J	50S ribosomal protein L3	1495557..1496180	624

*These columns are provided according to the reference genome *L. salivarius* UCC118

Table S5. Distribution of LPXTG-containing and sortase enzmyes across the 213 genomes

Species Name	StrainID	Proteins with LPXTG motif	LPXTG gene IDs	Sortase enzyme count	Sortase gene IDs	Initial LOCP output	Pilus-positive strain*	Number of clusters	Locus tags
Kandleria vitulina	DSM-20405	4	DSM_20405GL000656, DSM_20405GL000651, DSM_20405GL	7	DSM_20405GL000648, DSM_20405GL000647, D	Yes	1	1	DSM_20405GL000645-DSM_20405GL000657
Lactobacillus kimchiensis	DSM-24716	6	DSM_24716GL002188, DSM_24716GL000302, DSM_24716GL	1	DSM_24716GL0002255	No	0	0	
Lactobacillus mindensis	DSM-14500	9	DSM_14500GL000481, DSM_14500GL001981, DSM_14500GL	1	DSM_14500GL0001396	No	0	0	
Lactobacillus nantensis	DSM-16982	8	DSM_16982GL002497, DSM_16982GL001450, DSM_16982GL	1	DSM_16982GL002198	No	0	0	
Lactobacillus crustorum	LMG-23699	6	LMG_23699GL001827, LMG_23699GL000313, LMG_23699GL	1	LMG_23699GL001530	No	0	0	
Lactobacillus crustorum	JCM-15951	6	JCM_15951GL000895, JCM_15951GL000541, JCM_15951GL0	1	JCM_15951GL001441	No	0	0	
Lactobacillus futsaii	JCM-17355	5	JCM_17355GL001149, JCM_17355GL000512, JCM_17355GL0	1	JCM_17355GL001208	No	0	0	
Lactobacillus farciminis	DSM-20184	5	DSM_20184GL000284, DSM_20184GL001314, DSM_20184GL	1	DSM_20184GL001465	No	0	0	
Lactobacillus alimentarius	DSM-20249	5	DSM_20249GL001951, DSM_20249GL000453, DSM_20249GL	1	DSM_20249GL002165	No	0	0	
Lactobacillus paralimentarius	DSM-19674	6	DSM_19674GL002499, DSM_19674GL002023, DSM_19674GL	1	DSM_19674GL001187	No	0	0	
Lactobacillus paralimentarius	DSM-13961	6	DSM_13961GL000684, DSM_13961GL002440, DSM_13961GL	1	DSM_13961GL0021182	No	0	0	
Lactobacillus paralimentarius	DSM-13238	5	DSM_13238GL001227, DSM_13238GL000164, DSM_13238GL	1	DSM_13238GL000661	No	0	0	
Lactobacillus tucceti	DSM-20183	6	DSM_20183GL001473, DSM_20183GL000618, DSM_20183GL	1	DSM_20183GL000199	No	0	0	
Lactobacillus nodensis	DSM-19682	7	DSM_19682GL001072, DSM_19682GL000295, DSM_19682GL	1	DSM_19682GL000261	No	0	0	
Lactobacillus versmoldensis	DSM-14857	4	DSM_14857GL001286, DSM_14857GL002013, DSM_14857GL	1	DSM_14857GL000145	No	0	0	
Lactobacillus floricola	DSM-23037	0	0	1	DSM_23037GL000896	No	0	0	
Lactobacillus acidophilus	DSM-20079	11	DSM_20079GL000490, DSM_20079GL001512, DSM_20079GL	1	DSM_20079GL000934	No	0	0	
Lactobacillus amylovorus	DSM-20531	10	DSM_20531GL000807, DSM_20531GL001919, DSM_20531GL	1	DSM_20531GL000577	No	0	0	
Lactobacillus amylovorus	DSM-16698	6	DSM_16698GL001813, DSM_16698GL000171, DSM_16698GL	1	DSM_16698GL001620	No	0	0	
Lactobacillus kitasatonis	DSM-16761	11	DSM_16761GL001138, DSM_16761GL001485, DSM_16761GL	1	DSM_16761GL001604	No	0	0	
Lactobacillus kefiranofaciens kefirano	DSM-5016	8	DSM_5016GL001593, DSM_5016GL001467, DSM_5016GL000	2	DSM_5016GL000350, DSM_5016GL000099	No	0	0	
Lactobacillus kefiranofaciens kefirgra	DSM-10550	4	DSM_10550GL001210, DSM_10550GL002011, DSM_10550GL	1	DSM_10550GL000062	No	0	0	
Lactobacillus ultunensis	DSM-16047	3	DSM_16047GL001570, DSM_16047GL001218, DSM_16047GL	1	DSM_16047GL000409	No	0	0	
Lactobacillus helveticus	LMG-22464	9	LMG_22464GL001772, LMG_22464GL001551, LMG_22464GL	1	LMG_22464GL000006	No	0	0	
Lactobacillus helveticus	CGMCC-1.1877	5	CGMCC_1.1877GL000568, CGMCC_1.1877GL000883, CGMC	1	CGMCC_1.1877GL000871	No	0	0	
Lactobacillus gallinarum	DSM-10532	11	DSM_10532GL001210, DSM_10532GL000798, DSM_10532GL	1	DSM_10532GL001659	No	0	0	
Lactobacillus crispatus	DSM-20584	11	DSM_20584GL000238, DSM_20584GL000824, DSM_20584GL	1	DSM_20584GL001854	No	0	0	
Lactobacillus acetotolerans	DSM-20749	0	0	1	DSM_20749GL001333	Yes	0	0	
Lactobacillus intestinalis	DSM-6629	13	DSM_6629GL000820, DSM_6629GL001159, DSM_6629GL001	1	DSM_6629GL0000077	Yes	0	0	
Lactobacillus hamsteri	DSM-5661	10	DSM_5661GL000059, DSM_5661GL001728, DSM_5661GL000	1	DSM_5661GL000439	No	0	0	
Lactobacillus amylolyticus	DSM-11664	2	DSM_11664GL001247, DSM_11664GL000217	1	DSM_11664GL000186	No	0	0	
Lactobacillus kalixensis	DSM-16043	10	DSM_16043GL001402, DSM_16043GL000893, DSM_16043GL	1	DSM_16043GL001759	No	0	0	
Lactobacillus gigeriorum	DSM-23908	9	DSM_23908GL001420, DSM_23908GL000570, DSM_23908GL	1	DSM_23908GL000896	No	0	0	
Lactobacillus pasteurii	DSM-23907	11	DSM_23907GL001563, DSM_23907GL001030, DSM_23907GL	1	DSM_23907GL000159	No	0	0	
Lactobacillus delbrueckii jakobsenii	DSM-26046	1	DSM_26046GL001229	2	DSM_26046GL000643, DSM_26046GL000289	No	0	0	
Lactobacillus delbrueckii lactis	DSM-20072	2	DSM_20072GL001164, DSM_20072GL000182	2	DSM_20072GL000714, DSM_20072GL001764	No	0	0	
Lactobacillus delbrueckii bulgaricus	DSM-20081	2	DSM_20081GL001481, DSM_20081GL001900	1	DSM_20081GL001567	No	0	0	
Lactobacillus delbrueckii delbrueckii	DSM-20074	2	DSM_20074GL000853, _DSM_20074GL001664	1	DSM_20074GL001418	No	0	0	
Lactobacillus delbrueckii indicus	DSM-15996	1	DSM_15996GL000211	2	DSM_15996GL000941, DSM_15996GL001548	No	0	0	
Lactobacillus equicursoris	DSM-19284	0	0	3	DSM_19284GL000558, DSM_19284GL000191, D	Yes	1	1	DSM_19284GL000189-DSM_19284GL000193
Lactobacillus jensenii	DSM-20557	10	DSM_20557GL000947, DSM_20557GL000697, DSM_20557GL	1	DSM_20557GL000120	No	0	0	
Lactobacillus psittaci	DSM-15354	14	DSM_15354GL001320, DSM_15354GL001152, DSM_15354GL	1	DSM_15354GL0002286	Yes	0	0	
Lactobacillus hominis	DSM-23910	16	DSM_23910GL001758, DSM_23910GL000242, DSM_23910GL	1	DSM_23910GL000185	Yes	0	0	
Lactobacillus taiwanensis	DSM-21401	8	DSM_21401GL000754, DSM_21401GL001759, DSM_21401GL	1	DSM_21401GL0004422	No	0	0	
Lactobacillus johnsonii	ATCC-33200	8	ATCC_33200GL000689, ATCC_33200GL001379, ATCC_3320	1	ATCC_33200GL000369	No	0	0	
Lactobacillus gasseri	ATCC-33323	9	ATCC_3332300450, ATCC_3332300166, ATCC_3332301827, /	1	ATCC_3332300912	No	0	0	
Lactobacillus iners	DSM-13335	6	DSM_13335GL000092, DSM_13335GL000001, DSM_13335GL	1	DSM_13335GL0003007	No	0	0	
Lactobacillus amylotrophicus	DSM-20534	6	DSM_20534GL001366, DSM_20534GL000953, DSM_20534GL	4	DSM_20534GL001368, DSM_20534GL001384, D	Yes	1	1	DSM_20534GL000078-DSM_20534GL000080
Lactobacillus amylophilus	DSM-20533	6	DSM_20533GL000200, DSM_20533GL000760, DSM_20533GL	4	DSM_20533GL000713, DSM_20533GL001214, D	Yes	1	1	DSM_20533GL000759-DSM_20533GL000761
Lactobacillus dextrinicus	DSM-20335	4	DSM_20335GL000025, DSM_20335GL000538, DSM_20335GL	1	DSM_20335GL001519	No	0	0	
Lactobacillus concavus	DSM-17758	4	DSM_17758GL000815, DSM_17758GL000814, DSM_17758GL	2	DSM_17758GL0008118, DSM_17758GL000939	Yes	1	1	DSM_17758GL000814-DSM_17758GL000818
Lactobacillus composti	DSM-18527	6	DSM_18527GL002680, DSM_18527GL002347, DSM_18527GL	2	DSM_18527GL000267, DSM_18527GL002269	Yes	1	1	DSM_18527GL000264-DSM_18527GL000267
Lactobacillus harbinensis	DSM-16991	5	DSM_16991GL002809, DSM_16991GL000067, DSM_16991GL	7	DSM_16991GL000066, DSM_16991GL000067, D	Yes	1	1	DSM_16991GL002440-DSM_16991GL002444
Lactobacillus perolens	DSM-12744	11	DSM_12744GL002691, DSM_12744GL000201, DSM_12744GL	3	DSM_12744GL003091, DSM_12744GL003090, D	Yes	1	1	DSM_12744GL003087-DSM_12744GL003091
Lactobacillus camelliae	DSM-22697	16	DSM_22697GL001249, DSM_22697GL001930, DSM_22697GL	2	DSM_22697GL001227, DSM_22697GL001529	Yes	1	1	DSM_22697GL001224-DSM_22697GL001227
Lactobacillus nasuensis	JCM-17158	20	JCM_17158GL002113, JCM_17158GL001381, JCM_17158GL0	5	JCM_17158GL000214, JCM_17158GL001503, JC	Yes	1	3	JCM_17158GL001379-JCM_17158GL001384,JCM_17158GL001503-JCI
Lactobacillus manihotivorans	DSM-13343	9	DSM_13343GL001150, DSM_13343GL002810, DSM_13343GL	7	DSM_13343GL001490, DSM_13343GL001151, D	Yes	1	4	DSM_13343GL000818-DSM_13343GL000822, DSM_13343GL001392-D
Lactobacillus paracasei paracasei	DSM-5622	9	DSM_5622GL002008, DSM_5622GL002113, DSM_5622GL001	4	DSM_5622GL000483, DSM_5622GL001345, DSI	Yes	1	3	DSM_5622GL002077-DSM_5622GL002080, DSM_5622GL002703-DSI
Lactobacillus casei	DSM-20011	11	DSM_20011GL001051, DSM_20011GL000676, DSM_20011GL	4	DSM_20011GL000675, DSM_20011GL001014, D	Yes	1	2	DSM_20011GL000675-DSM_20011GL000679 & DSM_20011GL001884
Lactobacillus paracasei tolerans	DSM-20258	6	DSM_20258GL001272, DSM_20258GL000196, DSM_20258GL	3	DSM_20258GL000472, DSM_20258GL000385, D	Yes	1	1	DSM_20258GL001269-DSM_20258GL001273
Lactobacillus zeae	DSM-20178	15	DSM_20178GL001213, DSM_20178GL002422, DSM_20178GL	1	DSM_20178GL000621	Yes	0	0	
Lactobacillus rhamnosus	DSM-20021	11	DSM_20021GL001756, DSM_20021GL000043, DSM_20021GL	2	DSM_20021GL002528, DSM_20021GL000983	Yes	1	1	DSM_20021GL000983-DSM_20021GL000986
Lactobacillus saniviri	DSM-24301	4	DSM_24301GL001627, DSM_24301GL002208, DSM_24301GL	2	DSM_24301GL000206, DSM_24301GL000570	No	0	0	
Lactobacillus brantae	DSM-23927	13	DSM_23927GL001495, DSM_23927GL001391, DSM_23927GL	5	DSM_23927GL001737, DSM_23927GL001396, D	Yes	1	1	DSM_23927GL001391-DSM_23927GL001396
Lactobacillus thailandensis	DSM-22698	11	DSM_22698GL000980, DSM_22698GL001891, DSM_22698GL	6	DSM_22698GL000953, DSM_22698GL000950, D	Yes	1	2	DSM_22698GL000947-DSM_22698GL000951, DSM_22698GL001891-D
Lactobacillus pantheris	DSM-15945	13	DSM_15945GL002217, DSM_15945GL001936, DSM_15945GL	5	DSM_15945GL001147, DSM_15945GL001146, D	Yes	1	2	DSM_15945GL000317-DSM_15945GL000320, DSM_15945GL001143-D
Lactobacillus sharpeae	DSM-20505	17	DSM_20505GL000259, DSM_20505GL001026, DSM_20505GL	5	DSM_20505GL001290, DSM_20505GL000365, D	Yes	1	5	DSM_20505GL001771-DSM_20505GL001774, DSM_20505GL001023-D
Lactobacillus selangorensis	DSM-13344	6	DSM_13344GL000084, DSM_13344GL001050, DSM_13344GL	4	DSM_13344GL001052, DSM_13344GL000915, D	Yes	1	1	DSM_13344GL001049-DSM_13344GL001052
Lactobacillus selangorensis	ATCC-BAA-66	6	ATCC_BAA-66GL001894, ATCC_BAA-66GL000803, ATCC_J	4	ATCC_BAA-66GL001277, ATCC_BAA-66GL00	Yes	1	1	ATCC_BAA-66GL000801-ATCC_BAA-66GL000804
Lactobacillus graminis	DSM-20719	4	DSM_20719GL001158, DSM_20719GL000066, DSM_20719GL	2	DSM_20719GL000709, DSM_20719GL000591	No	0	0	
Lactobacillus curvatus	DSM-20019	2	DSM_20019GL001659, DSM_20019GL000479	1	DSM_20019GL000931	No	0	0	
Lactobacillus sakei sakei	DSM-20017	3	DSM_20017GL000885, DSM_20017GL001848, DSM_20017GL	2	DSM_20017GL001603, DSM_20017GL001759	No	0	0	
Lactobacillus sakei carnosus	DSM-15831	2	DSM_15831GL001730, DSM_15831GL001597	2	DSM_15831GL001007, DSM_15831GL000783	No	0	0	
Lactobacillus fuchuensis	DSM-14340	11	DSM_14340GL001061, DSM_14340GL000533, DSM_14340GL	2	DSM_14340GL001064, DSM_14340GL001839	No	0	0	
Lactobacillus rennini	DSM-20253	1	DSM_20253GL001773	1	DSM_20253GL001147	No	0	0	
Lactobacillus coryniformis torquens	DSM-20004	6	DSM_20004GL001254, DSM_20004GL001644, DSM_20004GL	1	DSM_20004GL001323	No	0	0	
Lactobacillus coryniformis coryniform	DSM-20001	3	DSM_20001GL001561, DSM_20001GL001914, DSM_20001GL	1	DSM_20001GL002353	No	0	0	
Lactobacillus bifermentans	DSM-20003	19	DSM_20003GL003041, DSM_20003GL001394, DSM_20003GL	1	DSM_20003GL000268	No	0	0	
Lactobacillus ceti	DSM-22408	12	DSM_22408GL001179, DSM_22408GL000237, DSM_22408GL	1	DSM_22408GL000079	Yes	0	0	
Lactobacillus saerimneri	DSM-16049	11	DSM_16049GL000506, DSM_16049GL000767, DSM_16049GL	1	DSM_16049GL000237	No	0	0	
Lactobacillus animalis	DSM-20602	7	DSM_20602GL001492, DSM_20602GL000660, DSM_20602GL	2	DSM_20602GL000309, DSM_20602GL000659	Yes	1	1	DSM_20602GL000659-DSM_20602GL000662
Lactobacillus murinus	DSM-20452	14	DSM_20452GL001615, DSM_20452GL001141, DSM_20452GL	4	DSM_20452GL000752, DSM_20452GL001251, D	Yes	1	2	DSM_20452GL001251-DSM_20452GL001254, DSM_20452GL001778-D
Lactobacillus apodemi	DSM-16634	17	DSM_16634GL001785, DSM_16634GL000537, DSM_16634GL	2	DSM_16634GL000825, DSM_16634GL000295	Yes	1	1	DSM_16634GL000822-DSM_16634GL000825
Lactobacillus ruminis	DSM-20403	2	DSM_20403GL001548, DSM_20403GL001549	2	DSM_20403GL001547, DSM_20403GL001202	Yes	1	1	DSM_20403GL001547-DSM_20403GL001550
Lactobacillus agilis	DSM-20509	1	DSM_20509GL000672	1	DSM_20509GL000133	No	0	0	
Lactobacillus equi	DSM-15833	10	DSM_15833GL000904, DSM_15833GL001458, DSM_15833GL	1	DSM_15833GL000512	No	0	0	
Lactobacillus salivarius	DSM-20555	10	DSM_20555GL001355, DSM_20555GL001345, DSM_20555GL	3	DSM_20555GL000340, DSM_20555GL001342, D	Yes	1	1	DSM_20555GL001342-DSM_20555GL001346
Lactobacillus hayakitensis	DSM-18933	12	DSM_18933GL001209, DSM_18933GL001131, DSM_18933GL	1	DSM_18933GL001461	No	0	0	
Lactobacillus pobuzihii	NBRC-103219	1	NBRC_103219GL001356	1	NBRC_103219GL001355	Yes	1	1	NBRC_103219GL001355-NBRC_103219GL001360
Lactobacillus pobuzihii.Chen	KCTC-13174	1	KCTC_13174GL000405	1	KCTC_13174GL000406	Yes	1	1	KCTC_13174GL000401-KCTC_13174GL000406
Lactobacillus acidipiscis	DSM-15836	0	0	0	0	Yes	0	0	
Lactobacillus acidipiscis	DSM-15353	0	0	0	0	No	0	0	
Lactobacillus aviarius aviarius	DSM-20655	15	DSM_20655GL000800, DSM_20655GL000912, DSM_20655GL	2	DSM_20655GL001022, DSM_20655GL000842	Yes	0	0	
Lactobacillus aviarius araffinosus	DSM-20653	4	DSM_20653GL000101, DSM_20653GL000406, DSM_20653GL	3	DSM_20653GL000808, DSM_20653GL001336, D	No	0	0	
Lactobacillus sucicola	DSM-21376	4	DSM_21376GL002225, DSM_21376GL000824, DSM_21376GL	1	DSM_21376GL000924	No	0	0	
Lactobacillus aquaticus	DSM-21051	5	DSM_21051GL001805, DSM_21051GL002236, DSM_21051GL	1	DSM_21051GL001874	No	0	0	
Lactobacillus uvarum	DSM-19971	4	DSM_19971GL001019, DSM_19971GL002388, DSM_19971GL	1	DSM_19971GL002229	No	0	0	
Lactobacillus capillatus	DSM-19910	2	DSM_19910GL002045, DSM_19910GL001092	1	DSM_19910GL001950	No	0	0	
Lactobacillus cacaonum	DSM-21116	3	DSM_21116GL000555, DSM_21116GL001136, DSM_21116GL	1	DSM_21116GL001133	Yes	0	0	
Lactobacillus mali	DSM-20444	4	DSM_20444GL002356, DSM_20444GL001903, DSM_20444GL	1	DSM_20444GL002197	No	0	0	
Lactobacillus mali	ATCC-27304	2	ATCC_27304GL000830, ATCC_27304GL000316	1	ATCC_27304GL000170	No	0	0	
Lactobacillus hordei	DSM-19519	2	DSM_19519GL000851, DSM_19519GL000949	1	DSM_19519GL001591	No	0	0	
Lactobacillus oeni	DSM-19972	1	DSM_19972GL000478	1	DSM_19972GL000450	No	0	0	
Lactobacillus satsumensis	DSM-16230	3	DSM_16230GL001686, DSM_16230GL001897, DSM_16230GL	1	DSM_16230GL001163	No	0	0	
Lactobacillus vini	DSM-20605	1	DSM_20605GL001645	1	DSM_20605GL001364	No	0	0	
Lactobacillus ghanensis	DSM-18630	5	DSM_18630GL000918, DSM_18630GL000355, DSM_18630GL	1	DSM_18630GL001725	No	0	0	
Lactobacillus nagelii	DSM-13675	3	DSM_13675GL000470, DSM_13675GL001778, DSM_13675GL	1	DSM_13675GL001651	No	0	0	
Lactobacillus algidus	DSM-15638	7	DSM_15638GL001423, DSM_15638GL000902, DSM_15638GL	1	DSM_15638GL001287	No	0	0	
Lactobacillus fabifermentans	DSM-21115	14	DSM_21115GL001291, DSM_21115GL000394, DSM_21115GL	1	DSM_21115GL000343	No	0	0	
Lactobacillus xiangfangensis	LMG-26013	17	LMG_26013GL002218, LMG_26013GL001093, LMG_26013GL	1	LMG_26013GL002749	No	0	0	
Lactobacillus pentosus	DSM-20314	17	DSM_20314GL000260, DSM_20314GL002586, DSM_20314GL	1	DSM_20314GL002810	No	0	0	
Lactobacillus plantarum argentoratens	DSM-16365	16	DSM_16365GL001172, DSM_16365GL000907, DSM_16365GL	1	DSM_16365GL000737	No	0	0	
Lactobacillus plantarum	DSM-13273	17	DSM_13273GL003164, DSM_13273GL003307, DSM_13273GL	1	DSM_13273GL002714	No	0	0	
Lactobacillus plantarum plantarum	CGMCC-1.2437	17	CGMCC_1.2437GL001848, CGMCC_1.2437GL002767, CGMC	1	CGMCC_1.2437GL000951	No	0	0	
Lactobacillus paraplantarum	DSM-10667	12	DSM_10667GL000604, DSM_10667GL003005, DSM_10667GL	2	DSM_10667GL002231, DSM_10667GL001666	No	0	0	
Lactobacillus siliginis	DSM-22696	4	DSM_22696GL000946, DSM_22696GL000633, DSM_22696GL	1	DSM_22696GL001637	Yes	0	0	
Lactobacillus rossiae	DSM-15814	12	DSM_15814GL002690, DSM_15814GL000659, DSM_15814GL	1	DSM_15814GL002440	No	0	0	
Weissella viridescens	DSM-20410	7	DSM_20410GL000919, DSM_20410GL001336, DSM_20410GL	1	DSM_20410GL000457	No	0	0	
Weissella minor	DSM-20014	6	DSM_20014GL001262, DSM_20014GL001263, DSM_20014GL	1	DSM_20014GL000152	No	0	0	
Weissella halotolerans	DSM-20190	1	DSM_20190GL000797	1	DSM_20190GL000080	No	0	0	
Weissella confusa	DSM-20196	2	DSM_20196GL002068, DSM_20196GL000011	3	DSM_20196GL000684, DSM_20196GL001573, D	Yes	1	1	DSM_20196GL000684-DSM_20196GL000687
Weissella kandleri	DSM-20593	1	DSM_20593GL000407	0	0	No	0	0	
Oenococcus oeni	ATCC-BAA-116	0	0	0	0	No	0	0	
Oenococcus kitaharae	DSM-17330	4	DSM_17330GL00021, DSM_17330GL02268, DSM_17330GL0	2	DSM_17330GL00019, DSM_17330GL00020	Yes	1	1	DSM_17330GL000019-DSM_17330GL000025
Leuconostoc fallax	KCTC-3537	1	KCTC_3537GL01532	0	0	No	0	0	
Leuconostoc pseudomesenteroides	4882	2	4882GL00444, 4882GL01575	0	0	No	0	0	
Leuconostoc mesenteroides	ATCC-8293	1	ATCC_8293GL01689	2	ATCC_8293GL01687, ATCC_8293GL01686	Yes	1	1	ATCC_8293GL001686-ATCC_8293GL001691
Leuconostoc mesenteroides cremoris	ATCC-19254	0	0	1	ATCC_19254GL00587	No	0	0	
Leuconostoc carnosum	JB16	3	JB16GL01760, JB16GL00061, JB16GL01763	2	JB16GL01758, JB16GL01759	Yes	1	1	JB16GL001758-JB16GL001763
Leuconostoc argentinum	KCTC-3773	2	KCTC_3773GL00587, KCTC_3773GL01754	1	KCTC_3773GL01755	No	0	0	
Leuconostoc citreum	KM20	1	KM20GL00082	0	0	No	0	0	
Leuconostoc gelidum	KCTC-3527	1	KCTC_3527GL00516	0	0	No	0	0	
Leuconostoc gasicomitatum	LMG-18811	0	0	0	0	No	0	0	
Leuconostoc kimchii	IMSNU-11154	3	IMSNU_11154GL01197, IMSNU_11154GL02005, IMSNU_111	2	IMSNU_11154GL01825, IMSNU_11154GL01824	Yes	1	1	IMSNU_11154GL001821-IMSNU_11154GL001825
Fructobacillus fructosus	DSM-20349	0	0	0	0	No	0	0	
Lactobacillus pontis	DSM-8475	7	DSM_8475GL000768, DSM_8475GL000331, DSM_8475GL00	1	DSM_8475GL001343	No	0	0	
Lactobacillus panis	DSM-26035	3	DSM_26035GL001091, DSM_26035GL001738, DSM_26035GL001	1	DSM_26035GL000046	No	0	0	
Lactobacillus oris	DSM-4864	17	DSM_4864GL001046, DSM_4864GL001819, DSM_4864GL001	1	DSM_4864GL000568	No	0	0	
Lactobacillus antri	DSM-16041	26	DSM_16041GL001256, DSM_16041GL000013, DSM_16041GL	2	DSM_16041GL001094, DSM_16041GL000799	Yes	0	0	
Lactobacillus reuteri	DSM-20016	5	DSM_20016GL000247, DSM_20016GL001914, DSM_20016GL	1	DSM_20016GL001231	No	0	0	
Lactobacillus vaginalis	DSM-5837	5	DSM_5837GL000330, DSM_5837GL000015, DSM_5837GL00	1	DSM_5837GL001172	No	0	0	
Lactobacillus frumenti	DSM-13145	3	DSM_13145GL000149, DSM_13145GL001023, DSM_13145GL	1	DSM_13145GL001525	No	0	0	
Lactobacillus fermentum	DSM-20055	5	DSM_20055GL000692, DSM_20055GL001065, DSM_20055GL	1	DSM_20055GL001680	No	0	0	
Lactobacillus equigenerosi	DSM-18793	5	DSM_18793GL001526, DSM_18793GL000962, DSM_18793GL	1	DSM_18793GL001354	No	0	0	
Lactobacillus gastricus	DSM-16045	3	DSM_16045GL000205, DSM_16045GL000773, DSM_16045GL	1	DSM_16045GL000048	No	0	0	
Lactobacillus ingluviei	DSM-15946	8	DSM_15946GL000381, DSM_15946GL000686, DSM_15946GL	1	DSM_15946GL000241	No	0	0	
Lactobacillus ingluviei	DSM-14792	14	DSM_14792GL001808, DSM_14792GL000615, DSM_14792GL	1	DSM_14792GL000402	No	0	0	
Lactobacillus secaliphilus	DSM-17896	6	DSM_17896GL000271, DSM_17896GL000056, DSM_17896GL	1	DSM_17896GL001397	No	0	0	
Lactobacillus coleohominis	DSM-14060	15	DSM_14060GL000186, DSM_14060GL001520, DSM_14060GL	1	DSM_14060GL001320	No	0	0	
Lactobacillus mucosae	DSM-13345	15	DSM_13345GL001360, DSM_13345GL002094, DSM_13345GL	1	DSM_13345GL000548	No	0	0	
Lactobacillus oligofermentans	DSM-15707	5	DSM_15707GL000001, DSM_15707GL001343, DSM_15707GL	2	DSM_15707GL000087, DSM_15707GL000044	No	0	0	
Lactobacillus hokkaidonensis	DSM-26202	13	DSM_26202GL001556, DSM_26202GL002130, DSM_26202GL	4	DSM_26202GL001535, DSM_26202GL001536, D	Yes	1	1	DSM_26202GL001535-DSM_26202GL001538
Lactobacillus suebicus	DSM-5007	15	DSM_5007GL000525, DSM_5007GL001404, DSM_5007GL000	1	DSM_5007GL002085	No	0	0	
Lactobacillus vaccinostercus	DSM-20634	7	DSM_20634GL001369, DSM_20634GL001300, DSM_20634GL	1	DSM_20634GL001201	No	0	0	
Lactobacillus parabrevis	LMG-11984	17	LMG_11984GL001839, LMG_11984GL000308, LMG_11984GL	3	LMG_11984GL002001, LMG_11984GL001996, L	Yes	1	1	LMG_11984GL002001-LMG_11984GL002004
Lactobacillus parabrevis	ATCC-53295	16	ATCC_53295GL000129, ATCC_53295GL000027, ATCC_5329	3	ATCC_53295GL001936, ATCC_53295GL002031	Yes	1	1	ATCC_53295GL002030-ATCC_53295GL002033
Lactobacillus hammesii	DSM-16381	18	DSM_16381GL002462, DSM_16381GL000826, DSM_16381GL	3	DSM_16381GL000002, DSM_16381GL000001, D	Yes	1	1	DSM_16381GL000001-DSM_16381GL000004
Lactobacillus paucivorans	DSM-22467	23	DSM_22467GL001771, DSM_22467GL000286, DSM_22467GL	3	DSM_22467GL000284, DSM_22467GL001653, D	Yes	1	1	DSM_22467GL001653-DSM_22467GL001656
Lactobacillus senmaizukei	DSM-21775	17	DSM_21775GL001681, DSM_21775GL001003, DSM_21775GL	3	DSM_21775GL001641, DSM_21775GL001140, D	Yes	1	1	DSM_21775GL001640-DSM_21775GL001643
Lactobacillus brevis	DSM-20054	7	DSM_20054GL001191, DSM_20054GL001910, DSM_20054GL	1	DSM_20054GL002361	Yes	0	0	

Lactobacillus koreensis	JCM-16448	9	JCM_16448GL000085, JCM_16448GL001438, JCM_16448GL0	3	JCM_16448GL001675, JCM_16448GL000224, JC	Yes	1	1	JCM_16448GL0000222-JCM_16448GL000225
Lactobacillus zymae	DSM-19395	16	DSM_19395GL002319, DSM_19395GL000381, DSM_19395GL	1	DSM_19395GL000934	No	0	0	
Lactobacillus acidifarinae	DSM-19394	14	DSM_19394GL001854, DSM_19394GL002248, DSM_19394GL	2	DSM_19394GL001800, DSM_19394GL000905	Yes	0	0	
Lactobacillus namurensis	DSM-19117	15	DSM_19117GL000223, DSM_19117GL002162, DSM_19117GL	1	DSM_19117GL001722	No	0	0	
Lactobacillus spicheri	DSM-15429	15	DSM_15429GL000237, DSM_15429GL001364, DSM_15429GL	1	DSM_15429GL000943	No	0	0	
Lactobacillus kimchicus	JCM-15530	7	JCM_15530GL002468, JCM_15530GL000731, JCM_15530GL0	1	JCM_15530GL000783	No	0	0	
Lactobacillus similis	DSM-23365	21	DSM_23365GL001270, DSM_23365GL000121, DSM_23365GL	1	DSM_23365GL001158	Yes	0	0	
Lactobacillus odoratitofui	DSM-19909	23	DSM_19909GL000375, DSM_19909GL000334, DSM_19909GL	1	DSM_19909GL002358	Yes	0	0	
Lactobacillus collinoides	DSM-20515	10	DSM_20515GL000349, DSM_20515GL002667, DSM_20515GL	1	DSM_20515GL000897	Yes	0	0	
Lactobacillus paracollinoides	DSM-15502	15	DSM_15502GL003136, DSM_15502GL000710, DSM_15502GL	1	DSM_15502GL001533	Yes	0	0	
Lactobacillus malefermentans	DSM-5705	12	DSM_5705GL001772, DSM_5705GL000489, DSM_5705GL000	1	DSM_5705GL000241	No	0	0	
Lactobacillus parabuchneri	DSM-5707	9	DSM_5707GL002343, DSM_5707GL002311, DSM_5707GL001	1	DSM_5707GL002384	No	0	0	
Lactobacillus parabuchneri	DSM-15352	11	DSM_15352GL001529, DSM_15352GL000299, DSM_15352GL	3	DSM_15352GL002182, DSM_15352GL001528, D	Yes	1	1	DSM_15352GL001527-DSM_15352GL001531
Lactobacillus buchneri	DSM-20057	5	DSM_20057GL002290, DSM_20057GL002208, DSM_20057GL	1	DSM_20057GL002253	No	0	0	
Lactobacillus otakiensis	DSM-19908	6	DSM_19908GL000970, DSM_19908GL000708, DSM_19908GL	1	DSM_19908GL002000	Yes	0	0	
Lactobacillus kefiri	DSM-20587	3	DSM_20587GL001017, DSM_20587GL001780, DSM_20587GL	1	DSM_20587GL001982	No	0	0	
Lactobacillus parakefiri	DSM-10551	6	DSM_10551GL001097, DSM_10551GL001348, DSM_10551GL	2	DSM_10551GL000796, DSM_10551GL004542	No	0	0	
Lactobacillus sunkii	DSM-19904	4	DSM_19904GL000561, DSM_19904GL002117, DSM_19904GL	1	DSM_19904GL000603	No	0	0	
Lactobacillus rapi	DSM-19907	6	DSM_19907GL001146, DSM_19907GL002649, DSM_19907GL	1	DSM_19907GL001524	No	0	0	
Lactobacillus kisonensis	DSM-19906	6	DSM_19906GL002469, DSM_19906GL002006, DSM_19906GL	1	DSM_19906GL001744	No	0	0	
Lactobacillus diolivorans	DSM-14421	4	DSM_14421GL000349, DSM_14421GL002640, DSM_14421GL	2	DSM_14421GL002029, DSM_14421GL002101	Yes	0	0	
Lactobacillus hilgardii	DSM-20176	2	DSM_20176GL001813, DSM_20176GL000973	1	DSM_20176GL000386	No	0	0	
Lactobacillus farraginis	DSM-18382	1	DSM_18382GL000336	1	DSM_18382GL002163	Yes	0	0	
Lactobacillus parafarraginis	DSM-18390	7	DSM_18390GL001600, DSM_18390GL000527, DSM_18390GL	2	DSM_18390GL002766, DSM_18390GL001314	No	0	0	
Lactobacillus senioris	DSM-24302	3	DSM_24302GL001476, DSM_24302GL001193, DSM_24302GL	3	DSM_24302GL001196, DSM_24302GL001052, D	Yes	1	1	DSM_24302GL001193-DSM_24302GL001196
Lactobacillus ozensis	DSM-23829	0	0	1	DSM_23829GL001214	No	0	0	
Lactobacillus kunkeei	DSM-12361	6	DSM_12361GL001115, DSM_12361GL001310, DSM_12361GL	1	DSM_12361GL001048	No	0	0	
Lactobacillus florum	DSM-22689	6	DSM_22689GL000154, DSM_22689GL000095, DSM_22689GL	1	DSM_22689GL000108	No	0	0	
Lactobacillus lindneri	DSM-20690	6	DSM_20690GL001120, DSM_20690GL001342, DSM_20690GL	1	DSM_20690GL000658	No	0	0	
Lactobacillus sanfranciscensis	DSM-20451	4	DSM_20451GL000977, DSM_20451GL000262, DSM_20451GL	1	DSM_20451GL000270	No	0	0	
Lactobacillus fructivorans	ATCC-27394	9	ATCC_27394GL000693, ATCC_27394GL000680, ATCC_2739	1	ATCC_27394GL000683	Yes	0	0	
Lactobacillus homohiochii	DSM-20571	9	DSM_20571GL000572, DSM_20571GL000664, DSM_20571GL	1	DSM_20571GL000680	Yes	0	0	
Lactobacillus fructivorans	DSM-20350	8	DSM_20350GL000774, DSM_20350GL000329, DSM_20350GL	1	DSM_20350GL001071	Yes	0	0	
Lactobacillus fructivorans	DSM-20203	8	DSM_20203GL001276, DSM_20203GL000494, DSM_20203GL	1	DSM_20203GL000504	Yes	0	0	
Pediococcus argentinicus	DSM-23026	4	DSM_23026GL001175, DSM_23026GL000894, DSM_23026GL	1	DSM_23026GL000528	No	0	0	
Pediococcus claussenii	DSM-14800	3	DSM_14800GL000637, DSM_14800GL001642, DSM_14800GL	2	DSM_14800GL001036, DSM_14800GL001113	Yes	1	1	DSM_14800GL001113-DSM_14800GL001115
Pediococcus pentosaceus	DSM-20336	3	DSM_20336GL000909, DSM_20336GL000913, DSM_20336GL	1	DSM_20336GL000314	No	0	0	
Pediococcus stilesii	DSM-18001	0	0	1	DSM_18001GL000326	No	0	0	
Pediococcus lolii	DSM-19927	3	DSM_19927GL001047, DSM_19927GL001900, DSM_19927GL	1	DSM_19927GL001675	Yes	0	0	
Pediococcus acidilactici	AS1-2696	3	AS1.2696GL001049, AS1.2696GL001783, AS1.2696GL000415	1	AS1.2696GL000951	No	0	0	
Pediococcus ethanolidurans	DSM-22301	9	DSM_22301GL000128, DSM_22301GL000049, DSM_22301GL	4	DSM_22301GL001131, DSM_22301GL000126, D	Yes	1	1	DSM_22301GL000125-DSM_22301GL000128
Pediococcus cellicola	DSM-17757	7	DSM_17757GL000421, DSM_17757GL001525, DSM_17757GL	2	DSM_17757GL001526, DSM_17757GL001909	No	0	0	
Pediococcus damnosus	DSM-20331	10	DSM_20331GL001090, DSM_20331GL001685, DSM_20331GL	3	DSM_20331GL001092, DSM_20331GL001093, D	Yes	1	1	DSM_20331GL001090-DSM_20331GL001093
Pediococcus inopinatus	DSM-20285	10	DSM_20285GL000864, DSM_20285GL000210, DSM_20285GL	4	DSM_20285GL001776, DSM_20285GL000520, D	Yes	1	1	DSM_20285GL000546-DSM_20285GL000549
Pediococcus parvulus	DSM-20332	10	DSM_20332GL002774, DSM_20332GL003420, DSM_20332GL	6	DSM_20332GL003316, DSM_20332GL001496, D	Yes	1	2	DSM_20332GL003313-DSM_20332GL003316, DSM_20332GL003417-D
Carnobacterium maltaromaticum	DSM-20730	22	DSM_20730GL003263, DSM_20730GL000714, DSM_20730GL	4	DSM_20730GL003099, DSM_20730GL003366, D	Yes	0	0	
Carnobacterium maltaromaticum	DSM-20342	27	DSM_20342GL002886, DSM_20342GL001571, DSM_20342GL	7	DSM_20342GL001625, DSM_20342GL002378, D	Yes	1	1	DSM_20342GL002941-DSM_20342GL002944
Carnobacterium maltaromaticum	DSM-20722	22	DSM_20722GL002004, DSM_20722GL001429, DSM_20722GL	7	DSM_20722GL000754, DSM_20722GL001730, D	Yes	1	1	DSM_20722GL000754-DSM_20722GL000757
Carnobacterium divergens	DSM-20623	18	DSM_20623GL001163, DSM_20623GL000792, DSM_20623GL	4	DSM_20623GL001820, DSM_20623GL001147, D	No	0	0	
Lactococcus lactis	LMG-7760	7	LMG_7760GL001946, LMG_7760GL001608, LMG_7760GL00	2	LMG_7760GL001539, LMG_7760GL001947	Yes	1	1	LMG_7760GL0001946-LMG_7760GL001949
Atopobium minutum	DSM-20586	9	DSM_20586GL001354, DSM_20586GL000632, DSM_20586GL	1	DSM_20586GL001167	No	0	0	
Olsenella uli	DSM-7084	5	DSM_7084GL001434, DSM_7084GL000839, DSM_7084GL001	2	DSM_7084GL001746, DSM_7084GL001525	Yes	1	1	DSM_7084GL001746-DSM_7084GL001749
Atopobium rimae	DSM-7090	0	0	1	DSM_7090GL000687	No	0	0	
Total		**1628**		**357**			**51**	**67**	

Footnote: The gene sequences for all locus tags listed above are provided as supplementary datasets 3,4 and
***** Cells colored in light gray indicate strain harboring at least 1 pilus gene cluster with a similar gene order a *L. rhamnosus* strain GG.

Table S6. Distribution and abundance of cell envelope proteins and associated anchoring domains and motifs

			Cell anchor type		
Species name	**StrainID**	**CEPs**	**LPxTG**	**SLAP**	**T**
Kandleria vitulina	DSM-20405	0	0	0	0
Lactobacillus kimchiensis	DSM-24716	0	0	0	0
Lactobacillus mindensis	DSM-14500	0	0	0	0
Lactobacillus nantensis	DSM-16982	0	0	0	0
Lactobacillus crustorum	LMG-23699	0	0	0	0
Lactobacillus crustorum	JCM-15951	0	0	0	0
Lactobacillus futsaii	JCM-17355	0	0	0	0
Lactobacillus farciminis	DSM-20184	0	0	0	0
Lactobacillus alimentarius	DSM-20249	0	0	0	0
Lactobacillus paralimentarius	DSM-19674	0	0	0	0
Lactobacillus paralimentarius	DSM-13961	0	0	0	0
Lactobacillus paralimentarius	DSM-13238	0	0	0	0
Lactobacillus tucceti	DSM-20183	0	0	0	0
Lactobacillus nodensis	DSM-19682	0	0	0	0
Lactobacillus versmoldensis	DSM-14857	0	0	0	0
Lactobacillus floricola	DSM-23037	0	0	0	0
Lactobacillus acidophilus	DSM-20079	1	0	1	0
Lactobacillus amylovorus	DSM-20531	0	0	0	0
Lactobacillus amylovorus	DSM-16698	0	0	0	0
Lactobacillus kitasatonis	DSM-16761	0	0	0	0
Lactobacillus kefiranofaciens kefiranofaciens	DSM-5016	2	0	2	0
Lactobacillus kefiranofaciens kefirgranum	DSM-10550	0	0	0	0
Lactobacillus ultunensis	DSM-16047	1	0	1	0
Lactobacillus helveticus	LMG-22464	1	0	1	0
Lactobacillus helveticus	CGMCC-1.1877	1	0	1	0
Lactobacillus gallinarum	DSM-10532	1	0	1	0
Lactobacillus crispatus	DSM-20584	0	0	0	0
Lactobacillus acetotolerans	DSM-20749	0	0	0	0
Lactobacillus intestinalis	DSM-6629	1	0	1	0
Lactobacillus hamsteri	DSM-5661	0	0	0	0
Lactobacillus amylolyticus	DSM-11664	0	0	0	0
Lactobacillus kalixensis	DSM-16043	1	0	1	0
Lactobacillus gigeriorum	DSM-23908	1	0	1	0
Lactobacillus pasteurii	DSM-23907	0	0	0	0
Lactobacillus delbrueckii jakobsenii	DSM-26046	1	0	1	0
Lactobacillus delbrueckii lactis	DSM-20072	1	0	1	0
Lactobacillus delbrueckii bulgaricus	DSM-20081	1	0	1	0
Lactobacillus delbrueckii delbrueckii	DSM-20074	1	0	1	0
Lactobacillus delbrueckii indicus	DSM-15996	1	0	1	0
Lactobacillus equicursoris	DSM-19284	1	0	1	0
Lactobacillus jensenii	DSM-20557	0	0	0	0
Lactobacillus psittaci	DSM-15354	0	0	0	0
Lactobacillus hominis	DSM-23910	1	0	1	0
Lactobacillus taiwanensis	DSM-21401	1	1	0	0
Lactobacillus johnsonii	ATCC-33200	0	0	0	0
Lactobacillus gasseri	ATCC-33323	0	0	0	0
Lactobacillus iners	DSM-13335	0	0	0	0
Lactobacillus amylotrophicus	DSM-20534	0	0	0	0
Lactobacillus amylophilus	DSM-20533	0	0	0	0
Lactobacillus dextrinicus	DSM-20335	0	0	0	0
Lactobacillus concavus	DSM-17758	1	1	0	0
Lactobacillus composti	DSM-18527	0	0	0	0
Lactobacillus harbinensis	DSM-16991	0	0	0	0
Lactobacillus perolens	DSM-12744	1	0	0	1
Lactobacillus camelliae	DSM-22697	2	**2**	0	0
Lactobacillus nasuensis	JCM-17158	1	1	0	0
Lactobacillus manihotivorans	DSM-13343	1	1	0	0
Lactobacillus paracasei paracasei	DSM-5622	2	**2**	0	0
Lactobacillus casei	DSM-20011	1	1	0	0
Lactobacillus paracasei tolerans	DSM-20258	0	0	0	0
Lactobacillus zeae	DSM-20178	2	2	0	0
Lactobacillus rhamnosus	DSM-20021	2	**2**	0	0
Lactobacillus saniviri	DSM-24301	0	0	0	0
Lactobacillus brantae	DSM-23927	0	0	0	0
Lactobacillus thailandensis	DSM-22698	0	0	0	0
Lactobacillus pantheris	DSM-15945	1	1	0	0
Lactobacillus sharpeae	DSM-20505	0	0	0	0
Lactobacillus selangorensis	DSM-13344	0	0	0	0
Lactobacillus selangorensis	ATCC-BAA-66	0	0	0	0
Lactobacillus graminis	DSM-20719	0	0	0	0
Lactobacillus curvatus	DSM-20019	0	0	0	0
Lactobacillus sakei sakei	DSM-20017	0	0	0	0
Lactobacillus sakei carnosus	DSM-15831	0	0	0	0
Lactobacillus fuchuensis	DSM-14340	0	0	0	0
Lactobacillus rennini	DSM-20253	0	0	0	0
Lactobacillus coryniformis torquens	DSM-20004	0	0	0	0
Lactobacillus coryniformis coryniformis	DSM-20001	0	0	0	0
Lactobacillus bifermentans	DSM-20003	0	0	0	0
Lactobacillus ceti	DSM-22408	0	0	0	0
Lactobacillus saerimneri	DSM-16049	0	0	0	0
Lactobacillus animalis	DSM-20602	1	1	0	0
Lactobacillus murinus	DSM-20452	1	1	0	0
Lactobacillus apodemi	DSM-16634	1	1	0	0
Lactobacillus ruminis	DSM-20403	1	1	0	0
Lactobacillus agilis	DSM-20509	0	0	0	0
Lactobacillus equi	DSM-15833	1	1	0	0
Lactobacillus salivarius	DSM-20555	1	1	0	0
Lactobacillus hayakitensis	DSM-18933	0	0	0	0
Lactobacillus pobuzihii	NBRC-103219	0	0	0	0

Lactobacillus pobuzihii.Chen	KCTC-13174	0	0	0	0
Lactobacillus acidipiscis	DSM-15836	0	0	0	0
Lactobacillus acidipiscis	DSM-15353	0	0	0	0
Lactobacillus aviarius aviarius	DSM-20655	0	0	0	0
Lactobacillus aviarius araffinosus	DSM-20653	0	0	0	0
Lactobacillus sucicola	DSM-21376	0	0	0	0
Lactobacillus aquaticus	DSM-21051	0	0	0	0
Lactobacillus uvarum	DSM-19971	0	0	0	0
Lactobacillus capillatus	DSM-19910	0	0	0	0
Lactobacillus cacaonum	DSM-21116	0	0	0	0
Lactobacillus mali	DSM-20444	0	0	0	0
Lactobacillus mali	ATCC-27304	0	0	0	0
Lactobacillus hordei	DSM-19519	0	0	0	0
Lactobacillus oeni	DSM-19972	0	0	0	0
Lactobacillus satsumensis	DSM-16230	0	0	0	0
Lactobacillus vini	DSM-20605	0	0	0	0
Lactobacillus ghanensis	DSM-18630	0	0	0	0
Lactobacillus nagelii	DSM-13675	0	0	0	0
Lactobacillus algidus	DSM-15638	0	0	0	0
Lactobacillus fabifermentans	DSM-21115	0	0	0	0
Lactobacillus xiangfangensis	LMG-26013	0	0	0	0
Lactobacillus pentosus	DSM-20314	0	0	0	0
Lactobacillus plantarum argentoratensis	DSM-16365	0	0	0	0
Lactobacillus plantarum	DSM-13273	0	0	0	0
Lactobacillus plantarum plantarum	CGMCC-1.2437	0	0	0	0
Lactobacillus paraplantarum	DSM-10667	0	0	0	0
Lactobacillus siliginis	DSM-22696	0	0	0	0
Lactobacillus rossiae	DSM-15814	0	0	0	0
Weissella viridescens	DSM-20410	0	0	0	0
Weissella minor	DSM-20014	0	0	0	0
Weissella halotolerans	DSM-20190	0	0	0	0
Weissella confusa	DSM-20196	0	0	0	0
Weissella kandleri	DSM-20593	0	0	0	0
Oenococcus oeni	ATCC-BAA-1163	0	0	0	0
Oenococcus kitaharae	DSM-17330	0	0	0	0
Leuconostoc fallax	KCTC-3537	0	0	0	0
Leuconostoc pseudomesenteroides	4882-	1	0	0	1
Leuconostoc mesenteroides	ATCC-8293	0	0	0	0
Leuconostoc mesenteroides cremoris	ATCC-19254	0	0	0	0
Leuconostoc carnosum	JB16-	0	0	0	0
Leuconostoc argentinum	KCTC-3773	0	0	0	0
Leuconostoc citreum	KM20-	0	0	0	0
Leuconostoc gelidum	KCTC-3527	0	0	0	0
Leuconostoc gasicomitatum	LMG-18811	0	0	0	0
Leuconostoc kimchii	IMSNU-11154	0	0	0	0
Fructobacillus fructosus	DSM-20349	0	0	0	0
Lactobacillus pontis	DSM-8475	0	0	0	0
Lactobacillus panis	DSM-6035	0	0	0	0
Lactobacillus oris	DSM-4864	0	0	0	0
Lactobacillus antri	DSM-16041	0	0	0	0
Lactobacillus reuteri	DSM-20016	0	0	0	0
Lactobacillus vaginalis	DSM-5837	0	0	0	0
Lactobacillus frumenti	DSM-13145	0	0	0	0
Lactobacillus fermentum	DSM-20055	0	0	0	0
Lactobacillus equigenerosi	DSM-18793	0	0	0	0
Lactobacillus gastricus	DSM-16045	0	0	0	0
Lactobacillus ingluviei	DSM-15946	0	0	0	0
Lactobacillus ingluviei	DSM-14792	0	0	0	0
Lactobacillus secaliphilus	DSM-17896	0	0	0	0
Lactobacillus coleohominis	DSM-14060	0	0	0	0
Lactobacillus mucosae	DSM-13345	0	0	0	0
Lactobacillus oligofermentans	DSM-15707	0	0	0	0
Lactobacillus hokkaidonensis	DSM-26202	0	0	0	0
Lactobacillus suebicus	DSM-5007	0	0	0	0
Lactobacillus vaccinostercus	DSM-20634	0	0	0	0
Lactobacillus parabrevis	LMG-11984	0	0	0	0
Lactobacillus parabrevis	ATCC-53295	0	0	0	0
Lactobacillus hammesii	DSM-16381	0	0	0	0
Lactobacillus paucivorans	DSM-22467	0	0	0	0
Lactobacillus senmaizukei	DSM-21775	0	0	0	0
Lactobacillus brevis	DSM-20054	0	0	0	0
Lactobacillus koreensis	JCM-16448	0	0	0	0
Lactobacillus zymae	DSM-19395	0	0	0	0
Lactobacillus acidifarinae	DSM-19394	0	0	0	0
Lactobacillus namurensis	DSM-19117	0	0	0	0
Lactobacillus spicheri	DSM-15429	0	0	0	0
Lactobacillus kimchicus	JCM-15530	0	0	0	0
Lactobacillus similis	DSM-23365	0	0	0	0
Lactobacillus odoratitofui	DSM-19909	0	0	0	0
Lactobacillus collinoides	DSM-20515	1	1	0	0
Lactobacillus paracollinoides	DSM-15502	0	0	0	0
Lactobacillus malefermentans	DSM-5705	0	0	0	0
Lactobacillus parabuchneri	DSM-5707	1	0	0	1
Lactobacillus parabuchneri	DSM-15352	1	0	0	1
Lactobacillus buchneri	DSM-20057	1	0	0	1
Lactobacillus otakiensis	DSM-19908	1	0	0	1
Lactobacillus kefiri	DSM-20587	1	0	0	1
Lactobacillus parakefiri	DSM-10551	2	0	0	2
Lactobacillus sunkii	DSM-19904	1	0	0	1
Lactobacillus rapi	DSM-19907	0	0	0	0
Lactobacillus kisonensis	DSM-19906	0	0	0	0
Lactobacillus diolivorans	DSM-14421	2	1	0	1
Lactobacillus hilgardii	DSM-20176	0	0	0	0
Lactobacillus farraginis	DSM-18382	1	0	0	1

Lactobacillus parafarraginis	DSM-18390	1	0	0	1
Lactobacillus senioris	DSM-24302	0	0	0	0
Lactobacillus ozensis	DSM-23829	0	0	0	0
Lactobacillus kunkeei	DSM-12361	0	0	0	0
Lactobacillus florum	DSM-22689	0	0	0	0
Lactobacillus lindneri	DSM-20690	0	0	0	0
Lactobacillus sanfranciscensis	DSM-20451	0	0	0	0
Lactobacillus fructivorans	ATCC-27394	0	0	0	0
Lactobacillus homohiochii	DSM-20571	0	0	0	0
Lactobacillus fructivorans	DSM-20350	0	0	0	0
Lactobacillus fructivorans	DSM-20203	0	0	0	0
Pediococcus argentinicus	DSM-23026	0	0	0	0
Pediococcus claussenii	DSM-14800	0	0	0	0
Pediococcus pentosaceus	DSM-20336	0	0	0	0
Pediococcus stilesii	DSM-18001	0	0	0	0
Pediococcus lolii	DSM-19927	0	0	0	0
Pediococcus acidilactici	AS1-2696	0	0	0	0
Pediococcus ethanolidurans	DSM-22301	1	1	0	0
Pediococcus cellicola	DSM-17757	1	1	0	0
Pediococcus damnosus	DSM-20331	0	0	0	0
Pediococcus inopinatus	DSM-20285	0	0	0	0
Pediococcus parvulus	DSM-20332	0	0	0	0
Carnobacterium maltaromaticum	DSM-20730	1	1	0	0
Carnobacterium maltaromaticum	DSM-20342	1	1	0	0
Carnobacterium maltaromaticum	DSM-20722	1	1	0	0
Carnobacterium divergens	DSM-20623	1	1	0	0
Lactococcus lactis	LMG-7760	0	0	0	0
Atopobium minutum	DSM-20586	2	**2**	0	0
Olsenella uli	DSM-7084	0	0	0	0
Atopobium rimae	DSM-7090	0	0	0	0
Total		**60**	**30**	**17**	**13**

Table S7. CRISPR occurrence and diversity.

Species Name	StrainID	CRISPR Type	CRISPR Repeat Sequence	DR Length	No. of Spacers	Cas1	Cas3	Cas9	Cas10
Kandleria vitulina	DSM-20405	II	GTTTTAGAGTTGTGTTATTTTGAACAGATACAAAAC	36	43	Y		Y	
Lactobacillus kimchiensis	DSM-24716	Undefined	GTGTTCCCCATATACATGGGGATGATTCT	29	3				
Lactobacillus mindensis A	DSM-14500	II	GTTTTAGAAGTAAGTCATCTCAATTACTAAGAACC	35	25	Y		Y	
Lactobacillus mindensis B	DSM-14500	Undefined	GTGCTCCCCATAAACATGGGGATGATTCT	29	3				
Lactobacillus nantensis	DSM-16982	II	GTTTTTGTACCTTCAAAGATTTAGAAACAGTAAAAC	36	22	Y		Y	
Lactobacillus crustorum	LMG-23699	N/A	None found						
Lactobacillus crustorum	JCM-15951	N/A	None found						
Lactobacillus futsaii	JCM-17355	II	GTTTTTGTACCTTAAAGAATCTAGAAATAGTAAAAC	36	19	Y		Y	
Lactobacillus farciminis A	DSM-20184	Undefined	GTTTTTGTACCTTAAAGAATCTAGAAATAGTAAAAC	36	4				
Lactobacillus farciminis B	DSM-20184	II	GTTTTAGAAGTATGTCTTTCTATTTACTTAAGAACC	36	11	Y		Y	
Lactobacillus alimentarius	DSM-20249	N/A	None found						
Lactobacillus paralimentarius	DSM-19674	N/A	None found						
Lactobacillus paralimentarius	DSM-13961	N/A	None found						
Lactobacillus paralimentarius A	DSM-13238	I	GTACTCCCCATGTATATGGGGATGATTCC	29	11	Y	Y		
Lactobacillus paralimentarius B	DSM-13238	Undefined	GTGCTCCCCATATACATGGGGATGATTCT	29	15				
Lactobacillus tucceti	DSM-20183	II	GTTTTTGTACTCTTAAGGATTTAGTAATAGTAAAAC	36	15	Y		Y	
Lactobacillus nodensis A	DSM-19682	II	GTTTTAGAAGTACGTCATTTCATGTAGTTAAGAACC	36	9	Y		Y	
Lactobacillus nodensis B	DSM-19682	II	GTTTTAGTACTCTCAAGAATTTAGTAACAGTAAAAC	36	9	Y		Y	
Lactobacillus versmoldensis	DSM-14857	II	GTTTTAGATCTAAGTCATCTCAATTACTTAAGAACC	36	2	Y		Y	
Lactobacillus floricola A	DSM-23037	II	GTTTTAGAAGTATGTCAAATCAATAAGGTTAATACC	36	5	Y		Y	
Lactobacillus floricola B	DSM-23037	II	GTTTCAGAAGTATGTCATTTGAATAATGTTAGGACT	36	5	Y		Y	
Lactobacillus floricola C	DSM-23037	Undefined	CTTTTTCTCGCATACGCGAGAGTGATCC	28	46				
Lactobacillus acidophilus	DSM-20079	Undefined	ATTTTCTCCACGTATGTGGAGGTGATCC	28	31				
Lactobacillus amylovorus	DSM-20531	N/A	None found						
Lactobacillus amylovorus	DSM-16698	I	GTTTTTATTTAACTTAAGAGAAATGTAAAT	30	53	Y	Y		
Lactobacillus kitasatonis	DSM-16761	N/A	None found						
Lactobacillus kefiranofaciens kefiranofaciens A	DSM-5016	Undefined	GTGTTCTCCACGTATGTGGAGGT	23	5				
Lactobacillus kefiranofaciens kefiranofaciens B	DSM-5016	Undefined	GTGTTCTCCACGTATGTGGAGGTGATCCT	29	4				
Lactobacillus kefiranofaciens kefirgranum	DSM-10550	I	GTGTTCTCCACGTATGTGGAGGTGATCC	28	61	Y	Y		
Lactobacillus ultunensis	DSM-16047	N/A	None found						
Lactobacillus helveticus	LMG-22464	I	GTATTCTCCACGTATGTGGAGGTGATCC	28	26	Y	Y		
Lactobacillus helveticus	CGMCC-1.1877	I	GTTTTTATTTAACTTAAGAGAAATGTAAAG	30	41	Y	Y		
Lactobacillus gallinarum	DSM-10532	N/A	None found						
Lactobacillus crispatus	DSM-20584	Undefined	GTATTCTCCACGTGTGTGGAGGTGATCC	28	3				
Lactobacillus acetotolerans A	DSM-20749	I	GTATTCTCCACGTATGTGGAGGTGATCCT	29	45	Y	Y		
Lactobacillus acetotolerans B	DSM-20749	Undefined	GTTTTAGATGATTGTTAGATCAATGAGGTTTAGACC	36	9				
Lactobacillus intestinalis A	DSM-6629	I	GTATTCCCCACATGGGTGGGGGTGATCC	28	7	Y	Y		
Lactobacillus intestinalis B	DSM-6629	I	GTATTCCCCACACACGTGGGGGTGATCC	28	13	Y	Y		
Lactobacillus intestinalis C	DSM-6629	I	GTTTTTATTTAACTTAAGAGGAATGTAAAT	30	9	Y*	Y*		
Lactobacillus intestinalis D	DSM-6629	Undefined	GGATCACCTCCACATACGTGGAGAACAC	28	11				
Lactobacillus hamsteri	DSM-5661	I	GTATTCTCCACGCATGTGGAGGTGATCC	28	43	Y	Y		
Lactobacillus amylolyticus	DSM-11664	N/A	None found						
Lactobacillus kalixensis	DSM-16043	II	GTTTTAGATCTGATGTACCTAGTTAAGATGTAAAAC	36	4	Y*			
Lactobacillus gigeriorum	DSM-23908	N/A	None found						
Lactobacillus pasteurii	DSM-23907	I	GTGGTCCCCACGTAAGTGGGGGTGATCC	28	9	Y	Y		
Lactobacillus delbrueckii jakobsenii	DSM-26046	II	GTTTTAGAAGGTTGTCTATTCAATAAGGTTTAACCC	36	11	Y		Y	
Lactobacillus delbrueckii lactis	DSM-20072	Undefined	None found			Y	Y		
Lactobacillus delbrueckii bulgaricus	DSM-20081	I	GTATTCCCCACGCAAGTGGGGGTGATCC	28	40	Y	Y		
Lactobacillus delbrueckii delbrueckii	DSM-20074	I	GTATTCCCCACGCAAGTGGGGGTGATCC	28	40	Y	Y		
Lactobacillus delbrueckii indicus	DSM-15996	I	GTATTCCCCACGCAAGTGGGGGTGATCC	28	39	Y	Y		
Lactobacillus equicursoris	DSM-19284	I	GTATTCCCCTCGTATGAGGGGGTGATCC	28	21	Y	Y		
Lactobacillus jensenii	DSM-20557	Undefined	None found			Y		Y	
Lactobacillus psittaci	DSM-15354	II	GTTTTAGAAGGTTGTTAAATCAGTAAGTTGAAAAAC	36	42	Y		Y	
Lactobacillus hominis	DSM-23910	II	GTTTTAGTTAGTTGTTAGATCAATAAGGTTTAGATC	36	9	Y		Y	
Lactobacillus taiwanensis	DSM-21401	N/A	None found						
Lactobacillus johnsonii	ATCC-33200	N/A	None found						
Lactobacillus gasseri	ATCC-33323	N/A	None found						
Lactobacillus iners	DSM-13335	N/A	None found						
Lactobacillus amylotrophicus	DSM-20534	Undefined	GTTTTCCCCGCACAGGCGGGGGTGATCC	28	52				
Lactobacillus amylophilus	DSM-20533	I	GTTTTCCCCGCACAGGCGGGGGTGATCC	28	52	Y	Y		
Lactobacillus dextrinicus	DSM-20335	I	GTTTTTATCTAAAGAGTATTGAATGTAAAT	30	100	Y	Y		
Lactobacillus concavus	DSM-17758	II	GTTTTAGAAGAGTGTCATTTCAATAAGGTTAAGATC	36	68	Y		Y*	
Lactobacillus composti A	DSM-18527	II	GTTCTGGAAGTATGTCAGATCAATGGATTCAAGAGC	36	10	Y		Y	
Lactobacillus composti B	DSM-18527	Undefined	CGCAACTCCTGTAGTTGCGTGAATTGAAAT	30	5				
Lactobacillus harbinensis	DSM-16991	II	GTCCCATTTTAGCCGATTCTGGAAGGATCCAATAGC	36	60	Y*		Y	
Lactobacillus perolens A	DSM-12744	Undefined	TTAGGGGTGCGTGAATTGAAAG	22	9				
Lactobacillus perolens B	DSM-12744	I	GTCGCATCCCTCGGGGTGCGTGAATTGAAAG	31	9	Y	Y		
Lactobacillus camelliae	DSM-22697	N/A	None found						
Lactobacillus nasuensis	JCM-17158	N/A	None found						
Lactobacillus manihotivorans	DSM-13343	N/A	None found						
Lactobacillus paracasei paracasei	DSM-5622	N/A	None found						
Lactobacillus casei	DSM-20011	I	GTTTTCCCCGCACATGCGGGGGTGATCC	28	20	Y	Y*		
Lactobacillus paracasei tolerans	DSM-20258	N/A	None found						
Lactobacillus zeae	DSM-20178	I	GTCGCAGTCTACGTGACTGCGTGAATTGAAAT	32	86	Y	Y		
Lactobacillus rhamnosus	DSM-20021	N/A	None found						
Lactobacillus saniviri	DSM-24301	II	GTTTTAGTTGGATGTCAGATCAAATAGGTTAAGCAC	36	58	Y		Y	
Lactobacillus brantae	DSM-23927	I	GTATTCCCCGTGCATACGGGGGTGATCC	28	45	Y	Y		
Lactobacillus thailandensis	DSM-22698	I	GTGTTCCCCGCAGGTGCGGGGGTAATCC	28	73	Y	Y		
Lactobacillus pantheris	DSM-15945	N/A	None found						
Lactobacillus sharpeae	DSM-20505	N/A	None found						
Lactobacillus selangorensis	DSM-13344	I	GTTTTTATTTAAACAATATGGAATGTAAAT	30	33	Y	Y		
Lactobacillus selangorensis A	ATCC-BAA-66	I	GTTTTTATTTAAACAATATGGAATGTAAAT	30	37	Y	Y		
Lactobacillus selangorensis B	ATCC-BAA-66	Undefined	CTTTTTATTTAAACAATGTGGAATGTAAAT	30	8				
Lactobacillus graminis	DSM-20719	II	GTTTTAGAAGAGTATCAAATCAATGAGTAGTTCAAC	36	35	Y		Y	
Lactobacillus curvatus	DSM-20019	Undefined	GTTTTAGAAGAGTATCAAATCAATGAGTAGTTCAAC	36	6				
Lactobacillus sakei sakei	DSM-20017	N/A	None found						
Lactobacillus sakei carnosus	DSM-15831	II	GTTTTAGAAGAGTATCAAATCAATGAGTAGTTCAAC	36	21	Y		Y	
Lactobacillus fuchuensis	DSM-14340	II	GTTTTAGAAGAGTATCAAATCAATGAGTTCATCAAC	36	27	Y		Y	
Lactobacillus rennini A	DSM-20253	I	GTCGCACTCTATATGGGTGCGTGAATTGAAAT	32	24	Y	Y		
Lactobacillus rennini B	DSM-20253	II	GTTTTAGAAGAGTATCAAATCAATGAGTTTATCAAC	36	12	Y		Y	
Lactobacillus coryniformis torquens	DSM-20004	II	GCTATTGATTCCTTCAGTTTTCAGCTAAAATAGATC	36	5	Y		Y	
Lactobacillus coryniformis coryniformis	DSM-20001	II	GTTTTAGAAGAGTGTTAAATCAATGAGTTTAGAACC	36	30	Y		Y	
Lactobacillus bifermentans	DSM-20003	I	GTATTCCCCGCACCAGCGGGGGTGATCC	28	118	Y	Y		
Lactobacillus ceti	DSM-22408	II	GTTTAAGACCTATCGAGAACAACACACTTCTCAAAC	36	16	Y		Y	
Lactobacillus saerimneri	DSM-16049	II	GTTTTTGTACTCTGAAGAACTTAGTAATGGAATAAC	36	8	Y*		Y*	
Lactobacillus animalis	DSM-20602	II	GTTTTAGAGCTATGTTGTTTTGTATGACTCCAAAAC	36	51	Y		Y	
Lactobacillus murinus	DSM-20452	N/A	None found						
Lactobacillus apodemi	DSM-16634	II	GTTTTAGAGCTATGTAGTTTTGTATGACTCCAAAAC	36	13	Y		Y	
Lactobacillus ruminis A	DSM-20403	III	GTTTTCGTCTCCTTCACTCGGAGATAGGTAATTATC	36	13	Y			Y
Lactobacillus ruminis B	DSM-20403	I	ATTTCAACTCACGCCCCCTATACAGAGGGCGAC	33	29	Y	Y		
Lactobacillus agilis A	DSM-20509	I	CTTCTCCCCACACTAGTGGGGGTAATCC	28	35	Y	Y		
Lactobacillus agilis B	DSM-20509	II	GTCTCAGAAGTATGTTAAATCAATGATGTTTAGTAC	36	32	Y		Y	
Lactobacillus agilis C	DSM-20509	Undefined	GTTTACATTCCTCTTAAGTTAAATAGATTC	30	4				
Lactobacillus equi A	DSM-15833	I	GTGTTCCCTACGTATGTAGGGGTAATCC	28	13	Y*			
Lactobacillus equi B	DSM-15833	I	GTGTTCCCCGTGTGTACGGGGGTGATCC	28	34	Y	Y		
Lactobacillus equi C	DSM-15833	Undefined	GTATTCCCTACGTATGTAGGGGTG	24	14				
Lactobacillus salivarius	DSM-20555	III	GTTTTCGTCTCCTTCATTCGGAGATATGTTCTTATT	36	11	Y			Y
Lactobacillus hayakitensis	DSM-18933	N/A	None found						
Lactobacillus pobuzihii	NBRC-103219	N/A	None found						
Lactobacillus pobuzihii Chen	KCTC-13174	N/A	None found						
Lactobacillus acidipiscis	DSM-15836	N/A	None found						
Lactobacillus acidipiscis A	DSM-15353	I	GTATTCCCCACGCATGTGGGGGTGATCC	28	22	Y	Y		
Lactobacillus acidipiscis B	DSM-15353	III	GTTCTCGTCCCCTATTACCGGGGATCATCTAATACC	36	39	Y			Y
Lactobacillus aviarius aviarius	DSM-20655	N/A	None found						
Lactobacillus aviarius araffinosus	DSM-20653	N/A	None found						
Lactobacillus sucicola A	DSM-21376	I	GTATTCCCCGCGTATGCGGGGGTGATCC	28	16	Y	Y		
Lactobacillus sucicola B	DSM-21376	I	GTATTTATTTAACTTAAGAGGAATGTAAAT	30	41	Y	Y		
Lactobacillus aquaticus	DSM-21051	I	GTATTCCCCGCGCATGCGGGGGTGATCC	28	26	Y	Y		
Lactobacillus uvarum	DSM-19971	Undefined	GTTTTAGAAGAGTGTTAAATCAATGAGTTTAGAACC	36	6				
Lactobacillus capillatus	DSM-19910	I	GTATTCCCCGCGCATGCGGGGGTGATCC	28	44	Y	Y		
Lactobacillus cacaonum	DSM-21116	II	GTTTTTGTACTCTCAACATTTTCCTATCAGTAAAAC	36	36	Y		Y	
Lactobacillus mali	DSM-20444	N/A	None found						
Lactobacillus mali	ATCC-27304	II	GTTTTTGCACTCTCAACATTTTCCTATCAATAAAAC	36	21	Y		Y	
Lactobacillus hordei	DSM-19519	II	GTTTTTGTACTCTCAACATTTTCCTATCAGTAAAAC	36	26	Y		Y	
Lactobacillus oeni	DSM-19972	I	GTATTCCCCACGTATGTGGGGGTGATCC	28	26	Y	Y		
Lactobacillus satsumensis	DSM-16230	N/A	None found						
Lactobacillus vini	DSM-20605	Undefined	GTTTTTGTACTCTTAAGGATTCAGTAACGGTAAAAC	36	23	Y*			
Lactobacillus ghanensis	DSM-18630	Undefined	GATTTTATTTAACTTAAGAGGAATGTAAAT	30	3				
Lactobacillus nagelii	DSM-13675	I	GATTCTATTTAACTTAAGAGGAATGTAAAG	30	19	Y	Y		

Lactobacillus algidus	DSM-15638	N/A	None found					
Lactobacillus fabifermentans	DSM-21115	N/A	None found					
Lactobacillus xiangfangensis	LMG-26013	N/A	None found					
Lactobacillus pentosus A	DSM-20314	I	CTGTTCCCCGYGTATGCGGGGGTGATCC	28	23	Y	Y	
Lactobacillus pentosus B	DSM-20314	II	GTCTTGAATAGTAGTCATATCAAACAGGTTTAGAAC	36	10	Y		Y
Lactobacillus plantarum argentoratensis A	DSM-16365	I	CTGTTCCCCGTGTATGCGGGGGTGATCC	28	12	Y	Y	
Lactobacillus plantarum argentoratensis B	DSM-16365	Undefined	CTATTCCCCGTACATACGGGGGTGATCC	28	10			
Lactobacillus plantarum	DSM-13273	N/A	None found					
Lactobacillus plantarum plantarum	CGMCC-1.2437	N/A	None found					
Lactobacillus paraplantarum	DSM-10667	N/A	None found					
Lactobacillus siliginis	DSM-22696	N/A	None found					
Lactobacillus rossiae	DSM-15814	II	GTTTTAGATGTATGTCAGATCAATAGGGTTAAGAAC	36	14	Y		Y
Weissella viridescens	DSM-20410	N/A	None found					
Weissella minor	DSM-20014	Undefined	GCATCAGCAAGTGTGTCTGAATC	23	4		Y	
Weissella halotolerans	DSM-20190	II	GCTTTAGATGTATGTCAGATTAATGGGGTTTATTCC	36	8	Y		Y
Weissella confusa	DSM-20196	N/A	None found					
Weissella kandleri	DSM-20593	II	GCTTCATATCTCTGTCAAATTAATGAGTTGTTTAGC	36	15	Y		Y
Oenococcus oeni	ATCC-BAA-1163	N/A	None found					
Oenococcus kitaharae	DSM-17330	II	GCTTCAGATGTGTGTCAGATCAATGAGGTAGAACCC	36	57	Y		Y
Leuconostoc fallax	KCTC-3537	N/A	None found					
Leuconostoc pseudomesenteroides	4882	II	GTATAAAGCCCCATTGATCTGACATACATCTGAAGC	36	6	Y		Y
Leuconostoc mesenteroides	ATCC-8293	N/A	None found					
Leuconostoc mesenteroides cremoris	ATCC-19254	N/A	None found					
Leuconostoc carnosum	JB16	N/A	None found					
Leuconostoc argentinum	KCTC-3773	N/A	None found					
Leuconostoc citreum	KM20	N/A	None found					
Leuconostoc gelidum	KCTC-3527	II	GCTTCAGATGTGTGTCAGATCAATGAGGTTTAAACC	36	29	Y		Y
Leuconostoc gasicomitatum	LMG-18811	N/A	None found					
Leuconostoc kimchii	IMSNU-11154	N/A	None found					
Fructobacillus fructosus	DSM-20349	II	GCTTTAGATGTATGTCGGATTAATGGGGTTTCTTCC	36	8	Y		Y
Lactobacillus pontis	DSM-8475	N/A	None found					
Lactobacillus panis	DSM-6035	Undefined	GTATTTATCTAATAGAAGTGGAATGTAAAT	30	30			
Lactobacillus oris	DSM-4864	I	GTATTCCCCATGTACGTGGGGGTGATCC	28	18	Y	Y	
Lactobacillus antri	DSM-16041	I	GTATTCCCCATGTGTATGGGGGTGATCC	28	10	Y	Y	
Lactobacillus reuteri	DSM-20016	N/A	None found					
Lactobacillus vaginalis	DSM-5837	N/A	None found					
Lactobacillus frumenti	DSM-13145	N/A	None found					
Lactobacillus fermentum	DSM-20055	II	GTACTCGGAACTACTGATCTGACACTCATCCAAGAC	36	6	Y		Y
Lactobacillus equigenerosi	DSM-18793	N/A	None found					
Lactobacillus gastricus	DSM-16045	II	GTTTTAGAAGGGCTTCAACTCAATGACCTCAAGACC	36	17	Y		Y
Lactobacillus ingluviei A	DSM-15946	I	TCGCACTCTGTAATGGAGTGCGTGGATTGAAAT	33	43	Y	Y	
Lactobacillus ingluviei B	DSM-15946	I	GTCGCATCCGGATCTTCGGGTGCGCGGATTGAAAT	35	6	Y*	Y*	
Lactobacillus ingluviei C	DSM-15946	Undefined	CTCTTTCCCCCGTATAGGGGGCTGAACCT	29	5			
Lactobacillus ingluviei A	DSM-14792	I	ATCGCACTCTGTAATGGAGTGCGTGGATTGAAAT	34	9	Y	Y	
Lactobacillus ingluviei B	DSM-14792	Undefined	CTCTTTCCCCCGCATAGGGGGCTGAACC	28	11			
Lactobacillus ingluviei C	DSM-14792	Undefined	GTCGCACTCTGTAATGGAGTGCGTGGATTGAAAT	34	16			
Lactobacillus secaliphilus	DSM-17896	II	GTTTTAGATGTACTTCAGATCAATGATGTTTAAAT	35	3	N		Y
Lactobacillus coleohominis	DSM-14060	N/A	None found					
Lactobacillus mucosae A	DSM-13345	I	GTATTCCCCATGTATGTGGGGGTGATCCT	29	25	Y	Y	
Lactobacillus mucosae B	DSM-13345	I	GTCGCACTCTCTTGTAGGGTGCGTGGATTGAAAT	34	20	Y	N	
Lactobacillus oligofermentans	DSM-15707	II	GTTTTAGAAGAGTATCAAATCAATGAGTTCGAGACC	36	30	Y		Y
Lactobacillus hokkaidonensis	DSM-26202	N/A	None found					
Lactobacillus suebicus	DSM-5007	I	GTCGCACTCTTCGTGAGTGCGTGAATTGAAAT	32	12	Y	Y	
Lactobacillus vaccinostercus	DSM-20634	N/A	None found					
Lactobacillus parabrevis	LMG-11984	II	GTTTTAAGTAGATGGTAAATCAATAAGGTCAAGAGC	36	30	N		
Lactobacillus parabrevis	ATCC-53295	II	GTTTTAAGTAGATGGTAAATCAATAAGGTCAAGAGC	36	30	N		
Lactobacillus hammesii A	DSM-16381	I	GTATTCCCCATGTGTATGGGGATGATCC	28	24	Y	Y	
Lactobacillus hammesii B	DSM-16381	II	GCTTTAGATTAGTGACAATTCAGTAAGGTCAAGAGC	36	47	Y		Y
Lactobacillus paucivorans	DSM-22467	I	GTACTCCCCACGTATGTGGGGATGATCC	28	16	Y	Y	
Lactobacillus senmaizukei	DSM-21775	II	GTTTTAGTTCAGTGACAAATCAATAGAGTTAAGAAC	36	39	Y		Y
Lactobacillus brevis	DSM-20054	N/A	None found					
Lactobacillus koreensis	JCM-16448	I	GTATTCCCCGTGTATACGGGGGTGATCC	28	62	Y	Y	
Lactobacillus zymae A	DSM-19395	II	GTTTTAGATGAGTGGTAAATCTAGAAGGTCAAAAGC	36	24	Y		Y
Lactobacillus zymae B	DSM-19395	Undefined	GTATTCCCCACATGTGTGGGGGTGATT	28	45			
Lactobacillus acidifarinae A	DSM-19394	I	GTATTCCCCACACGTGTGGGGGTGATTC	28	121	Y	Y	
Lactobacillus acidifarinae B	DSM-19394	Undefined	GCTTTAAGTTCCTTTCAAATTAGGGTATAAATTAAG	36	16			
Lactobacillus namurensis A	DSM-19117	I	GTATTCCCCGCGTATGCGGGGGTGATTC	28	28	Y	Y	
Lactobacillus namurensis B	DSM-19117	II	GTTTCAAATGAGTGGTAAATCTATAAGGTCAATACC	36	16	Y		Y
Lactobacillus spicheri	DSM-15429	I	GTATTCCCCACACGTGTGGGGGTGATCC	28	40	Y	Y	
Lactobacillus kimchicus	JCM-15530	I	GTACTCTCCGCACAYGCGGAGGTGATCC	28	54	Y	Y	
Lactobacillus similis A	DSM-23365	I	GTCTCGCTCTTTCGTGGAGCGAGTGAATTGAAAT	34	22	N	Y	
Lactobacillus similis B	DSM-23365	Undefined	ATCGCACTCTCTAAAGGAGTGCGTGAATTGAAAT	34	18			
Lactobacillus odoratitofui	DSM-19909	I	GTATTCTCCGCGTATGCGGAGGTGATCC	28	135	Y	Y	
Lactobacillus collinoides	DSM-20515	I	GTATTCCCCGCATGTGCGGGGGTGATCC	28	42	Y	Y	
Lactobacillus paracollinoides	DSM-15502	II	GTTTTAGATGAATATCAAATCAATGAGGTTCAAAGC	36	11	Y		Y
Lactobacillus malefermentans	DSM-5705	I	GTCGCACCTTGTATGAGGTGCGTGAATTGAAAT	33	26	Y	Y	
Lactobacillus parabuchneri	DSM-5707	Undefined	GTATTCCCCACGCATGTGGGGGTAATCC	28	10			
Lactobacillus parabuchneri	DSM-15352	II	GTTTTAGAAGGATGTTAAATCAATAAGGTTAAACCC	36	15	Y*		
Lactobacillus buchneri	DSM-20057	II	GTTTTAGAAGGATGTTAAATCAATAAGGTTAAACCC	36	9	Y		Y
Lactobacillus otakiensis A	DSM-19908	I	GTATTCCCCACGTATGTGGGGGTGATCC	28	11	Y	Y	
Lactobacillus otakiensis B	DSM-19908	II	GTTTTAGAAGGATGTTAAATCAATAAGGTTAAACCC	36	44	Y		Y
Lactobacillus kefiri A	DSM-20587	I	GTATTCCCCACGCATGTGGGGGTGATCC	28	23		Y	
Lactobacillus kefiri B	DSM-20587	Undefined	GTTTTAGAAGGATGTTAAATCAATAAGGTTAAACCC	36	10			
Lactobacillus parakefiri A	DSM-10551	I	GTATTCCCCACGCATGTGGGGGTGATCC	28	13	Y	Y	
Lactobacillus parakefiri B	DSM-10551	II	GTTTTAGAAGGATGTTAAATCAATAAGGTTAAACCC	36	34	Y		Y
Lactobacillus sunkii	DSM-19904	N/A	None found					
Lactobacillus rapi	DSM-19907	I	GTATTCCCCACGTGTGTGGGGGTGATCC	28	37	Y	Y	
Lactobacillus kisonensis	DSM-19906	Undefined	GTATTCCCCACGTATGTAGGGGTGATCC	34	19			
Lactobacillus diolivorans	DSM-14421	II	GTTTTAGAAGACTGTTGAATCAATGATGTTTAGTCC	36	74	Y		Y
Lactobacillus hilgardii	DSM-20176	N/A	None found					
Lactobacillus farraginis	DSM-18382	I	GTATTCCCCACGCATGTGGGGGTGATCC	28	9	Y	Y	
Lactobacillus parafarraginis	DSM-18390	I	GTATTCCCCACGCATGTGGGGGTGATCC	28	126	Y	Y	
Lactobacillus senioris	DSM-24302	N/A	None found					
Lactobacillus ozensis	DSM-23829	II	GTTATTAATACATTACAAATACATAGATGTTATAAT	36	37	Y		Y
Lactobacillus kunkeei	DSM-12361	N/A	None found					
Lactobacillus florum	DSM-22689	II	GTTTTAGAAGTACGTCATTCTAATGAGATTAAGAAC	36	10			Y*
Lactobacillus lindneri A	DSM-20690	II	GTTCTTAATCTCATTAGAATGACGTACTTCTAAAAC	36	23	Y		Y
Lactobacillus lindneri B	DSM-20690	Undefined	GTTTTGTAAGTACGTCATTTCAAGTAGTATTAAACC	36	2			
Lactobacillus sanfranciscensis	DSM-20451	N/A	None found					
Lactobacillus fructivorans	ATCC-27394	N/A	None found					
Lactobacillus homohiochii	DSM-20571	N/A	None found					
Lactobacillus fructivorans	DSM-20350	N/A	None found					
Lactobacillus fructivorans	DSM-20203	N/A	None found					
Pediococcus argentinicus	DSM-23026	N/A	None found					
Pediococcus claussenii	DSM-14800	N/A	None found					
Pediococcus pentosaceus	DSM-20336	N/A	None found					
Pediococcus stilesii	DSM-18001	II	GTTTCAGAGGGATGTTAAAGAAGTAGGTTCAATATC	36	28	Y		Y
Pediococcus lolii	DSM-19927	II	GTTTCAGAAGGATGTTAAATCAATAAGGTTAAGATC	36	19	Y		Y
Pediococcus acidilactici	AS1.2696	II	GTTTCAGAAGGATGTTAAATCAATAAGGTTAAGATC	36	11	Y		Y*
Pediococcus ethanolidurans	DSM-22301	N/A	None found					
Pediococcus cellicola	DSM-17757	I	GTCACATCCTTTATGGATGTGTGAATTGAAAT	32	70	Y	Y	
Pediococcus damnosus	DSM-20331	II	GTTTTAGAAGAGTGTCGAATCAATATAGTTAAGAGC	36	20	Y		Y*
Pediococcus inopinatus	DSM-20285	II	GTTTTAGAAGAGTGTCGAATCAATATAGTTAAGATC	36	80	Y		Y
Pediococcus parvulus A	DSM-20332	II	GTTTCAGAAGGATGTTAAATCAATAAGGTTAAGATC	36	15	Y		Y
Pediococcus parvulus B	DSM-20332	II	GTTTTAGAAGAGTGTCGAATCAATATAGTTAAGAGC	36	90	Y		Y
Pediococcus parvulus C	DSM-20332	I	GTCGCATCCTTTATGGGTGCGTGAATTGAAAT	32	80	Y	Y	
Carnobacterium maltaromaticum	DSM-20730	N/A	None found					
Carnobacterium maltaromaticum	DSM-20342	N/A	None found					
Carnobacterium maltaromaticum	DSM-20722	N/A	None found					
Carnobacterium divergens	DSM-20623	N/A	None found					
Lactococcus lactis	LMG-7760	N/A	None found					
Atopobium minutum	DSM-20586	N/A	None found					
Olsenella uli	DSM-7084	II	GTTTTGGGGCAGTGTCGTTTTGACTGGTAATCAAAC	36	30	Y		Y
Atopobium rimae	DSM-7090	Undefined	GTTTTGGAGCAGTGTCATTCTGACTGGTAATCAAAC	36	5	Y*		

Table S8. Sequence information for the 27 core partial genes

PID*	Gene	*L. salivarius* locus tag*	COG	Annotation	Co-ordinates*	Strand*	Length*
90960991	*dnaA*	LSL_0001	COG0593L	chromosomal replication initiation protein	1..1365	+	454
90960994	*recF*	LSL_0004	COG1195L	recombination protein F	3309..4448	+	379
90961461	*murC*	LSL_0485	COG0773M	UDP-N-acetylmuramate--L-alanine ligase	533809..535140	+	443
90961551	*ribF*	LSL_0575	COG0196H	riboflavin kinase/FMN adenylyltransferase	616816..617772	+	318
90961553	*grpE*	LSL_0577	COG0576O	GrpE protein HSP-70 cofactor	618942..619538	+	198
90961554	*dnaK*	LSL_0578	COG0443O	molecular chaperone DnaK	619577..621424	+	615
90961556	*lepA*	LSL_0580	COG0481M	GTP-binding protein LepA	622843..624672	+	609
90961566	*prfA*	LSL_0590	COG0216J	peptide chain release factor 1	633672..634754	+	360
90961572	*atpF*	LSL_0596	COG0711C	ATP synthase subunit B	638577..639104	+	175
90961615	*rpsO*	LSL_0638	COG0184J	30S ribosomal protein S15	681489..681758	+	89
90962017	*ileS*	LSL_1042	COG0060J	isoleucyl-tRNA synthetase	1065203..1067998	-	931
90962030	*mraW*	LSL_1055	COG0275M	S-adenosyl-methyltransferase MraW	1080326..1081270	-	314
90962140	*pgk*	LSL_1165	COG0126G	phosphoglycerate kinase	1199234..1200436	-	400
90962149	*uvrA*	LSL_1174	COG0178L	excinuclease ABC subunit A	1208147..1210981	-	944
90962150	*uvrB*	LSL_1175	COG0556L	excinuclease ABC subunit B	1211000..1213000	-	666
90962161	*secA*	LSL_1186	COG0653U	preprotein translocase subunit SecA	1223594..1225957	-	787
90962203	*dnaX*	LSL_1228	COG2812L	DNA polymerase III subunit gamma/tau	1260545..1262284	-	579
90962211	*rplL*	LSL_1237	COG0222J	50S ribosomal protein L7/L12	1270871..1271239	-	122
90962242	*galU*	LSL_1268	COG1210M	UTP--glucose-1-phosphate uridylyltransferase	1304213..1305085	-	290
90962313	*trmA*	LSL_1341	COG2265J	tRNA (Uracil-5-) -methyltransferase	1392529..1393908	-	459
90962321	*pcrA*	LSL_1349	COG0210L	ATP-dependent DNA helicase	1403083..1405317	-	744
90962375	*truA*	LSL_1404	COG0101J	tRNA pseudouridine synthase A	1478890..1479660	-	256
90962386	*secY*	LSL_1415	COG0201U	preprotein translocase subunit SecY	1486063..1487361	-	432
90962403	*rplB*	LSL_1432	COG0090J	50S ribosomal protein L2	1493766..1494599	-	277
90962565	*parB*	LSL_1596	COG1475K	chromosome partitioning protein, DNA-binding protein	1676145..1677020	-	291
90962568	*gidB*	LSL_1599	COG0357M	16S rRNA methyltransferase GidB	1678668..1679393	-	241
90962695	*dnaB*	LSL_1726	COG0305L	replicative DNA helicase	1807560..1808951	-	463

*These columns are provided according to the reference genome *L. salivarius* UCC118

Table S9. Presence and absence of the complete pathways for production of the 20 standard amino acids.

Species Name	StrainID	Histidine	Valine	Leucine	Isoleucine	Threonine	Lysine	Asparagine	Aspartate	Alanine	Serine	Glycine	Methionine	Cysteine	Tryptophan	Phenlyalenine	Tyrosine	Arginine	Proline	Glutamate	Glutamine
Kandleria vitulina	DSM-20405	P	P	P	P	P	A	P	A	A	P	P	A	P	P	A	A	P	A	P	P
Lactobacillus kimchiensis	DSM-24716	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus mindensis	DSM-14500	A	A	A	A	A	P	A	A	A	P	A	A	A	P	A	A	P	A	P	P
Lactobacillus nantensis	DSM-16982	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	P	A	A	P
Lactobacillus crustorum	LMG-23699	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus crustorum	JCM-15951	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus futsaii	JCM-17355	P	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	P	A	P	P
Lactobacillus farciminis	DSM-20184	P	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	P	A	A	P
Lactobacillus alimentarius	DSM-20249	A	A	A	A	P	A	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus paralimentarius	DSM-19674	A	A	A	A	A	P	A	A	A	A	P	A	A	A	A	A	A	A	A	P
Lactobacillus paralimentarius	DSM-13961	A	A	A	A	A	P	P	A	A	P	P	A	A	A	A	A	A	A	A	A
Lactobacillus paralimentarius	DSM-13238	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus tucceti	DSM-20183	P	A	A	A	A	P	P	A	P	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus nodensis	DSM-19682	P	A	A	A	P	P	P	A	A	A	A	A	A	P	A	A	A	A	A	P
Lactobacillus versmoldensis	DSM-14857	P	P	P	P	P	P	P	P	A	A	A	A	A	P	A	A	P	A	P	P
Lactobacillus floricola	DSM-23037	A	A	A	A	P	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A
Lactobacillus acidophilus	DSM-20079	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus amylovorus	DSM-20531	A	A	A	A	P	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus amylovorus	DSM-16698	A	A	A	A	P	P	P	A	A	P	A	A	A	A	A	A	A	A	A	P
Lactobacillus kitasatonis	DSM-16761	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus kefiranofaciens kefiranofaciens	DSM-5016	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus kefiranofaciens kefirgranum	DSM-10550	A	A	A	A	A	P	A	A	A	P	A	P	A	A	A	A	A	A	A	P
Lactobacillus ultunensis	DSM-16047	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus helveticus	LMG-22464	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus helveticus	CGMCC-1.1877	A	A	A	A	A	A	P	P	P	A	A	A	P	A	A	A	A	A	A	A
Lactobacillus gallinarum	DSM-10532	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus crispatus	DSM-20584	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus acetotolerans	DSM-20749	P	P	P	P	A	P	A	A	A	A	A	P	A	A	A	A	P	A	P	P
Lactobacillus intestinalis	DSM-6629	A	A	A	A	A	P	P	P	P	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus hamsteri	DSM-5661	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus amylolyticus	DSM-11664	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus kalixensis	DSM-16043	A	A	A	A	A	P	P	P	A	P	A	A	A	A	A	A	A	A	A	P
Lactobacillus gigeriorum	DSM-23908	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus pasteurii	DSM-23907	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus delbrueckii jakobsenii	DSM-26046	A	A	A	A	P	P	P	A	A	A	A	P	A	A	A	A	A	A	A	P
Lactobacillus delbrueckii lactis	DSM-20072	A	A	A	A	P	P	P	A	A	A	A	P	A	A	A	A	A	A	A	P
Lactobacillus delbrueckii bulgaricus	DSM-20081	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus delbrueckii delbrueckii	DSM-20074	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus delbrueckii indicus	DSM-15996	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus equicursoris	DSM-19284	A	A	A	A	P	P	P	A	A	A	A	P	A	A	A	A	A	A	A	P
Lactobacillus jensenii	DSM-20557	P	A	A	A	P	P	A	A	A	A	A	P	A	P	A	A	P	A	A	P
Lactobacillus psittaci	DSM-15354	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus hominis	DSM-23910	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus taiwanensis	DSM-21401	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus johnsonii	ATCC-33200	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus gasseri	ATCC-33323	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus iners	DSM-13335	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus amylotrophicus	DSM-20534	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus amylophilus	DSM-20533	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus dextrinicus	DSM-20335	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus concavus	DSM-17758	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus composti	DSM-18527	P	P	P	P	P	P	A	A	A	A	P	P	A	A	P	A	A	P	A	P
Lactobacillus harbinensis	DSM-16991	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus perolens	DSM-12744	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus camelliae	DSM-22697	A	A	A	A	A	A	P	P	A	P	A	A	A	A	A	A	A	A	A	P
Lactobacillus nasuensis	JCM-17158	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus manihotivorans	DSM-13343	P	A	P	A	P	P	P	P	P	A	P	P	A	P	A	A	P	A	P	P
Lactobacillus paracasei paracasei	DSM-5622	P	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	P	P	P
Lactobacillus casei	DSM-20011	A	A	A	A	P	P	A	A	P	A	A	A	A	A	A	A	A	A	P	P
Lactobacillus paracasei tolerans	DSM-20258	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	P	P
Lactobacillus zeae	DSM-20178	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus rhamnosus	DSM-20021	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	P	P
Lactobacillus saniviri	DSM-24301	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus brantae	DSM-23927	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus thailandensis	DSM-22698	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus pantheris	DSM-15945	A	A	A	A	A	P	A	P	A	A	P	P	P	A	A	A	A	A	A	P
Lactobacillus sharpeae	DSM-20505	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	P	A	P
Lactobacillus selangorensis	DSM-13344	A	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus selangorensis	ATCC-BAA-66	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus graminis	DSM-20719	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus curvatus	DSM-20019	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus sakei sakei	DSM-20017	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus sakei carnosus	DSM-15831	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus fuchuensis	DSM-14340	A	A	A	A	A	A	A	P	A	P	A	A	A	A	A	A	A	A	A	P
Lactobacillus rennini	DSM-20253	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus coryniformis torquens	DSM-20004	P	P	P	P	P	P	P	A	A	P	P	P	P	A	A	A	P	A	P	P
Lactobacillus coryniformis coryniformis	DSM-20001	P	P	P	P	P	P	P	A	A	P	P	P	P	A	A	A	P	A	P	P
Lactobacillus bifermentans	DSM-20003	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P	P
Lactobacillus ceti	DSM-22408	A	A	A	A	A	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A
Lactobacillus saerimneri	DSM-16049	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus animalis	DSM-20602	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus murinus	DSM-20452	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus apodemi	DSM-16634	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus ruminis	DSM-20403	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	P	A	P
Lactobacillus agilis	DSM-20509	A	A	A	A	A	A	A	P	A	A	A	A	P	A	A	A	A	A	A	P
Lactobacillus equi	DSM-15833	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	P	A	P
Lactobacillus salivarius	DSM-20555	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus hayakitensis	DSM-18933	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus pobuzihii	NBRC-103219	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus pobuzihii.Chen	KCTC-13174	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus acidipiscis	DSM-15836	A	A	A	A	A	A	P	P	A	A	A	A	P	A	A	A	A	A	A	P
Lactobacillus acidipiscis	DSM-15353	A	A	A	A	A	A	P	P	A	A	A	A	P	P	A	A	A	A	A	A
Lactobacillus aviarius aviarius	DSM-20655	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus aviarius araffinosus	DSM-20653	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus sucicola	DSM-21376	P	P	P	P	P	P	P	A	A	A	A	A	P	A	P	A	A	A	P	P
Lactobacillus aquaticus	DSM-21051	P	P	P	P	P	P	P	A	A	A	A	A	P	A	P	A	A	P	A	P
Lactobacillus uvarum	DSM-19971	P	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P
Lactobacillus capillatus	DSM-19910	P	P	P	P	P	P	P	A	A	A	A	A	A	A	P	A	A	P	A	P
Lactobacillus cacaonum	DSM-21116	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	P	A	P
Lactobacillus mali	DSM-20444	P	P	P	P	P	A	P	A	A	A	A	A	P	A	P	A	A	P	P	P
Lactobacillus mali	ATCC-27304	P	P	P	P	P	P	P	A	A	A	A	P	P	A	P	P	A	P	P	P
Lactobacillus hordei	DSM-19519	P	A	P	A	A	P	P	P	A	A	A	A	A	P	A	A	A	A	P	A	P	P
Lactobacillus oeni	DSM-19972	P	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	P	A	P	P
Lactobacillus satsumensis	DSM-16230	P	P	P	P	P	P	P	A	A	A	P	P	A	P	P	A	A	A	P	A	P	P
Lactobacillus vini	DSM-20605	P	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P	A	P	A
Lactobacillus ghanensis	DSM-18630	P	P	P	P	P	P	P	A	A	A	P	A	A	A	P	A	A	A	P	A	P	P
Lactobacillus nagelii	DSM-13675	P	A	P	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	P	A	P	P
Lactobacillus algidus	DSM-15638	A	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus fabifermentans	DSM-21115	P	A	A	A	A	P	A	A	A	A	A	A	P	P	A	A	A	A	A	A	P
Lactobacillus xiangfangensis	LMG-26013	P	A	A	A	A	P	A	P	A	A	A	A	A	P	A	A	A	A	A	A	P
Lactobacillus pentosus	DSM-20314	P	A	P	A	P	A	P	A	A	A	P	P	P	P	P	A	A	A	A	P	P
Lactobacillus plantarum argentoratensis	DSM-16365	P	A	A	A	P	P	P	A	A	P	P	P	P	P	A	A	P	A	A	A	P
Lactobacillus plantarum	DSM-13273	P	A	A	A	A	A	P	P	A	A	P	A	P	P	P	A	P	P	A	A	A
Lactobacillus plantarum plantarum	CGMCC-1.2437	A	A	A	A	A	A	P	P	P	A	P	A	P	A	P	A	A	A	P	A	P
Lactobacillus paraplantarum	DSM-10667	P	A	A	A	A	A	A	P	P	A	P	A	P	P	P	A	A	P	P	P	P
Lactobacillus siliginis	DSM-22696	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus rossiae	DSM-15814	A	A	A	A	A	A	P	A	A	A	A	A	A	P	A	A	A	A	A	A	A
Weissella viridescens	DSM-20410	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Weissella minor	DSM-20014	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Weissella halotolerans	DSM-20190	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Weissella confusa	DSM-20196	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Weissella kandleri	DSM-20593	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Oenococcus oeni	ATCC-BAA-1163	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Oenococcus kitaharae	DSM-17330	P	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	P	A	A	A	P
Leuconostoc fallax	KCTC-3537	P	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	P	A	A	A	P
Leuconostoc pseudomesenteroides	4882	P	A	A	A	A	P	A	A	A	A	A	A	P	A	A	A	A	P	A	A	P
Leuconostoc mesenteroides	ATCC-8293	P	P	P	P	P	P	A	A	A	A	A	P	P	P	P	P	A	A	P	A	P	P
Leuconostoc mesenteroides cremoris	ATCC-19254	P	P	P	P	P	P	A	A	A	A	A	P	P	P	P	P	A	A	P	A	A	P
Leuconostoc carnosum	JB16	P	P	P	P	P	P	A	A	A	A	A	P	P	P	P	P	A	A	P	A	A	P
Leuconostoc argentinum	KCTC-3773	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Leuconostoc citreum	KM20	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	P	A	A	P
Leuconostoc gelidum	KCTC-3527	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	P	A	A	P	A	A	P
Leuconostoc gasicomitatum	LMG-18811	P	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	P	A	A	P
Leuconostoc kimchii	IMSNU-11154	P	A	A	A	A	P	A	A	A	A	A	A	P	A	P	A	A	A	P	A	A	P
Fructobacillus fructosus	DSM-20349	A	A	A	A	A	P	A	A	A	A	A	A	A	A	P	A	A	A	A	A	A	P
Lactobacillus pontis	DSM-8475	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	P	A	P
Lactobacillus panis	DSM-6035	A	A	A	A	A	A	P	P	P	A	A	A	P	A	A	A	A	A	A	P	P	P
Lactobacillus oris	DSM-4864	A	A	A	A	A	P	P	A	P	P	A	P	P	P	A	A	A	A	A	A	A	A
Lactobacillus antri	DSM-16041	A	A	A	A	A	A	P	P	A	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus reuteri	DSM-20016	A	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus vaginalis	DSM-5837	A	A	A	A	A	P	P	P	P	A	A	A	P	A	A	A	A	A	A	A	P	P
Lactobacillus frumenti	DSM-13145	A	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus fermentum	DSM-20055	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	P	A	A	A	P
Lactobacillus equigenerosi	DSM-18793	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus gastricus	DSM-16045	P	A	A	A	A	P	P	A	A	A	A	A	P	A	A	A	A	P	A	A	A	P
Lactobacillus ingluviei	DSM-15946	A	A	A	P	A	P	P	P	A	A	A	P	A	A	P	A	A	A	A	A	A	P
Lactobacillus ingluviei	DSM-14792	A	A	A	A	A	P	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus secaliphilus	DSM-17896	A	A	A	A	A	A	P	A	A	P	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus coleohominis	DSM-14060	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus mucosae	DSM-13345	A	A	A	A	A	P	P	P	A	A	A	A	P	A	A	A	A	P	A	A	A	P
Lactobacillus oligofermentans	DSM-15707	A	A	A	A	A	P	P	A	A	A	A	A	A	P	A	A	A	A	A	A	A	P
Lactobacillus hokkaidonensis	DSM-26202	P	A	A	A	A	P	A	A	A	P	A	A	A	P	P	A	A	P	A	A	A	P
Lactobacillus suebicus	DSM-5007	P	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	A	P	A	A	P
Lactobacillus vaccinostercus	DSM-20634	A	P	P	P	P	P	A	A	A	A	P	P	A	P	P	A	A	A	P	A	P	P
Lactobacillus parabrevis	LMG-11984	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus parabrevis	ATCC-53295	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus hammesii	DSM-16381	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus paucivorans	DSM-22467	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus senmaizukei	DSM-21775	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus brevis	DSM-20054	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Lactobacillus koreensis	JCM-16448	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus zymae	DSM-19395	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus acidifarinae	DSM-19394	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Lactobacillus namurensis	DSM-19117	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus spicheri	DSM-15429	A	A	A	A	A	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	P
Lactobacillus kimchicus	JCM-15530	P	P	P	P	P	A	P	P	A	P	P	P	P	P	P	A	A	P	A	P	P
Lactobacillus similis	DSM-23365	P	P	P	P	P	A	P	A	P	P	P	P	P	P	A	A	P	P	P	P	P
Lactobacillus odoratitofui	DSM-19909	P	A	A	A	P	P	P	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Lactobacillus collinoides	DSM-20515	P	P	P	P	P	P	P	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Lactobacillus paracollinoides	DSM-15502	P	P	P	P	P	P	P	A	A	P	P	P	A	P	A	A	P	A	P	P	P
Lactobacillus malefermentans	DSM-5705	A	A	A	A	A	P	A	A	A	A	A	A	A	P	A	A	A	A	A	P	P
Lactobacillus parabuchneri	DSM-5707	P	P	P	A	P	P	A	A	A	A	A	P	A	P	A	A	P	P	P	P	P
Lactobacillus parabuchneri	DSM-15352	P	A	P	A	P	P	A	A	A	A	A	P	A	P	A	A	P	A	P	P	P
Lactobacillus buchneri	DSM-20057	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus otakiensis	DSM-19908	P	A	A	A	P	P	A	A	P	A	A	P	A	P	A	A	P	A	A	P	P
Lactobacillus kefiri	DSM-20587	A	A	A	A	P	A	A	A	A	A	A	A	A	P	A	A	P	A	P	P	P
Lactobacillus parakefiri	DSM-10551	P	A	A	A	P	P	A	A	A	A	A	P	A	P	A	A	P	P	P	P	P
Lactobacillus sunkii	DSM-19904	P	A	A	A	P	P	A	A	A	A	A	P	A	P	A	A	P	A	A	P	P
Lactobacillus rapi	DSM-19907	P	P	P	P	P	P	A	A	A	P	A	A	P	P	A	A	P	A	P	P	P
Lactobacillus kisonensis	DSM-19906	A	A	A	A	P	P	A	A	A	A	A	A	P	P	A	A	A	A	A	P	P
Lactobacillus diolivorans	DSM-14421	P	A	A	A	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Lactobacillus hilgardii	DSM-20176	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Lactobacillus farraginis	DSM-18382	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus parafarraginis	DSM-18390	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Lactobacillus senioris	DSM-24302	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus ozensis	DSM-23829	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus kunkeei	DSM-12361	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	P	A	P	P
Lactobacillus florum	DSM-22689	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus lindneri	DSM-20690	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus sanfranciscensis	DSM-20451	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus fructivorans	ATCC-27394	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus homohiochii	DSM-20571	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus fructivorans	DSM-20350	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Lactobacillus fructivorans	DSM-20203	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus argentinicus	DSM-23026	A	A	A	A	A	P	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A
Pediococcus claussenii	DSM-14800	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Pediococcus pentosaceus	DSM-20336	A	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus stilesii	DSM-18001	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	P	A	A	P	A
Pediococcus lolii	DSM-19927	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus acidilactici	AS1-2696	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus ethanolidurans	DSM-22301	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus cellicola	DSM-17757	A	A	A	A	A	P	A	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Pediococcus damnosus	DSM-20331	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus inopinatus	DSM-20285	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Pediococcus parvulus	DSM-20332	A	A	A	A	A	P	P	A	A	A	A	A	A	A	A	A	A	A	A	A	P
Carnobacterium maltaromaticum	DSM-20730	A	P	P	P	P	A	P	A	A	A	A	A	P	P	A	A	P	A	A	P	P
Carnobacterium maltaromaticum	DSM-20342	A	P	P	P	A	A	P	A	A	A	A	P	P	P	A	A	P	A	A	A	P
Carnobacterium maltaromaticum	DSM-20722	A	P	P	P	P	A	A	A	A	A	A	A	A	P	A	A	P	A	A	A	P
Carnobacterium divergens	DSM-20623	A	P	P	P	P	P	P	A	A	P	P	A	P	A	A	A	P	A	A	A	A
Lactococcus lactis	LMG-7760	P	P	P	P	P	P	A	A	A	P	P	P	P	P	A	A	P	A	P	P	P
Atopobium minutum	DSM-20586	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Olsenella uli	DSM-7084	P	A	A	A	P	P	A	P	P	A	A	A	A	A	A	A	A	A	A	P	P
Atopobium rimae	DSM-7090	A	A	A	A	P	A	P	A	A	A	A	P	A	A	A	A	A	A	A	A	P

Table S10. Presence of sirtuin homologs in the 213 genomes analyzed

Species Name	StrainID	SIR2 family proteins (homolog of mammalian SIRT1, SIR2L1, Sir2α)
Kandleria vitulina	DSM-20405	2
Lactobacillus kimchiensis	DSM-24716	1
Lactobacillus mindensis	DSM-14500	1
Lactobacillus nantensis	DSM-16982	1
Lactobacillus crustorum	LMG-23699	1
Lactobacillus crustorum	JCM-15951	1
Lactobacillus futsaii	JCM-17355	3
Lactobacillus farciminis	DSM-20184	2
Lactobacillus alimentarius	DSM-20249	2
Lactobacillus paralimentarius	DSM-19674	2
Lactobacillus paralimentarius	DSM-13961	2
Lactobacillus paralimentarius	DSM-13238	1
Lactobacillus tucceti	DSM-20183	2
Lactobacillus nodensis	DSM-19682	1
Lactobacillus versmoldensis	DSM-14857	3
Lactobacillus floricola	DSM-23037	0
Lactobacillus acidophilus	DSM-20079	1
Lactobacillus amylovorus	DSM-20531	1
Lactobacillus amylovorus	DSM-16698	1
Lactobacillus kitasatonis	DSM-16761	1
Lactobacillus kefiranofaciens kefiranofaciens	DSM-5016	3
Lactobacillus kefiranofaciens kefirgranum	DSM-10550	3
Lactobacillus ultunensis	DSM-16047	4
Lactobacillus helveticus	LMG-22464	3
Lactobacillus helveticus	CGMCC-1.1877	3
Lactobacillus gallinarum	DSM-10532	1
Lactobacillus crispatus	DSM-20584	1
Lactobacillus acetotolerans	DSM-20749	1
Lactobacillus intestinalis	DSM-6629	1
Lactobacillus hamsteri	DSM-5661	3
Lactobacillus amylolyticus	DSM-11664	1
Lactobacillus kalixensis	DSM-16043	1
Lactobacillus gigeriorum	DSM-23908	1
Lactobacillus pasteurii	DSM-23907	1
Lactobacillus delbrueckii jakobsenii	DSM-26046	1
Lactobacillus delbrueckii lactis	DSM-20072	1
Lactobacillus delbrueckii bulgaricus	DSM-20081	1
Lactobacillus delbrueckii delbrueckii	DSM-20074	1
Lactobacillus delbrueckii indicus	DSM-15996	1
Lactobacillus equicursoris	DSM-19284	1
Lactobacillus jensenii	DSM-20557	1
Lactobacillus psittaci	DSM-15354	1
Lactobacillus hominis	DSM-23910	1
Lactobacillus taiwanensis	DSM-21401	1

Lactobacillus johnsonii	ATCC-33200	2
Lactobacillus gasseri	ATCC-33323	1
Lactobacillus iners	DSM-13335	0
Lactobacillus amylotrophicus	DSM-20534	1
Lactobacillus amylophilus	DSM-20533	1
Lactobacillus dextrinicus	DSM-20335	1
Lactobacillus concavus	DSM-17758	1
Lactobacillus composti	DSM-18527	1
Lactobacillus harbinensis	DSM-16991	1
Lactobacillus perolens	DSM-12744	1
Lactobacillus camelliae	DSM-22697	1
Lactobacillus nasuensis	JCM-17158	1
Lactobacillus manihotivorans	DSM-13343	1
Lactobacillus paracasei paracasei	DSM-5622	1
Lactobacillus casei	DSM-20011	1
Lactobacillus paracasei tolerans	DSM-20258	1
Lactobacillus zeae	DSM-20178	2
Lactobacillus rhamnosus	DSM-20021	2
Lactobacillus saniviri	DSM-24301	1
Lactobacillus brantae	DSM-23927	1
Lactobacillus thailandensis	DSM-22698	1
Lactobacillus pantheris	DSM-15945	1
Lactobacillus sharpeae	DSM-20505	0
Lactobacillus selangorensis	DSM-13344	1
Lactobacillus selangorensis	ATCC-BAA-66	1
Lactobacillus graminis	DSM-20719	3
Lactobacillus curvatus	DSM-20019	3
Lactobacillus sakei sakei	DSM-20017	0
Lactobacillus sakei carnosus	DSM-15831	0
Lactobacillus fuchuensis	DSM-14340	0
Lactobacillus rennini	DSM-20253	1
Lactobacillus coryniformis torquens	DSM-20004	2
Lactobacillus coryniformis coryniformis	DSM-20001	1
Lactobacillus bifermentans	DSM-20003	1
Lactobacillus ceti	DSM-22408	1
Lactobacillus saerimneri	DSM-16049	0
Lactobacillus animalis	DSM-20602	0
Lactobacillus murinus	DSM-20452	0
Lactobacillus apodemi	DSM-16634	1
Lactobacillus ruminis	DSM-20403	1
Lactobacillus agilis	DSM-20509	1
Lactobacillus equi	DSM-15833	2
Lactobacillus salivarius	DSM-20555	1
Lactobacillus hayakitensis	DSM-18933	0
Lactobacillus pobuzihii	NBRC-103219	1
Lactobacillus pobuzihii.Chen	KCTC-13174	1
Lactobacillus acidipiscis	DSM-15836	1

Lactobacillus acidipiscis	DSM-15353	1
Lactobacillus aviarius aviarius	DSM-20655	1
Lactobacillus aviarius araffinosus	DSM-20653	1
Lactobacillus sucicola	DSM-21376	1
Lactobacillus aquaticus	DSM-21051	1
Lactobacillus uvarum	DSM-19971	1
Lactobacillus capillatus	DSM-19910	1
Lactobacillus cacaonum	DSM-21116	0
Lactobacillus mali	DSM-20444	0
Lactobacillus mali	ATCC-27304	0
Lactobacillus hordei	DSM-19519	0
Lactobacillus oeni	DSM-19972	1
Lactobacillus satsumensis	DSM-16230	1
Lactobacillus vini	DSM-20605	1
Lactobacillus ghanensis	DSM-18630	1
Lactobacillus nagelii	DSM-13675	1
Lactobacillus algidus	DSM-15638	0
Lactobacillus fabifermentans	DSM-21115	1
Lactobacillus xiangfangensis	LMG-26013	1
Lactobacillus pentosus	DSM-20314	1
Lactobacillus plantarum argentoratensis	DSM-16365	1
Lactobacillus plantarum	DSM-13273	1
Lactobacillus plantarum plantarum	CGMCC-1.2437	1
Lactobacillus paraplantarum	DSM-10667	1
Lactobacillus siliginis	DSM-22696	1
Lactobacillus rossiae	DSM-15814	1
Weissella viridescens	DSM-20410	0
Weissella minor	DSM-20014	0
Weissella halotolerans	DSM-20190	1
Weissella confusa	DSM-20196	1
Weissella kandleri	DSM-20593	0
Oenococcus oeni	ATCC-BAA-11(	1
Oenococcus kitaharae	DSM-17330	1
Leuconostoc fallax	KCTC-3537	0
Leuconostoc pseudomesenteroides	4882	1
Leuconostoc mesenteroides	ATCC-8293	1
Leuconostoc mesenteroides cremoris	ATCC-19254	1
Leuconostoc carnosum	JB16	0
Leuconostoc argentinum	KCTC-3773	1
Leuconostoc citreum	KM20	1
Leuconostoc gelidum	KCTC-3527	0
Leuconostoc gasicomitatum	LMG-18811	0
Leuconostoc kimchii	IMSNU-11154	0
Fructobacillus fructosus	DSM-20349	1
Lactobacillus pontis	DSM-8475	1
Lactobacillus panis	DSM-6035	2
Lactobacillus oris	DSM-4864	1

Lactobacillus antri	DSM-16041	1
Lactobacillus reuteri	DSM-20016	1
Lactobacillus vaginalis	DSM-5837	1
Lactobacillus frumenti	DSM-13145	2
Lactobacillus fermentum	DSM-20055	1
Lactobacillus equigenerosi	DSM-18793	1
Lactobacillus gastricus	DSM-16045	1
Lactobacillus ingluviei	DSM-15946	1
Lactobacillus ingluviei	DSM-14792	1
Lactobacillus secaliphilus	DSM-17896	1
Lactobacillus coleohominis	DSM-14060	0
Lactobacillus mucosae	DSM-13345	1
Lactobacillus oligofermentans	DSM-15707	1
Lactobacillus hokkaidonensis	DSM-26202	1
Lactobacillus suebicus	DSM-5007	1
Lactobacillus vaccinostercus	DSM-20634	1
Lactobacillus parabrevis	LMG-11984	1
Lactobacillus parabrevis	ATCC-53295	1
Lactobacillus hammesii	DSM-16381	1
Lactobacillus paucivorans	DSM-22467	1
Lactobacillus senmaizukei	DSM-21775	1
Lactobacillus brevis	DSM-20054	1
Lactobacillus koreensis	JCM-16448	1
Lactobacillus zymae	DSM-19395	1
Lactobacillus acidifarinae	DSM-19394	1
Lactobacillus namurensis	DSM-19117	1
Lactobacillus spicheri	DSM-15429	1
Lactobacillus kimchicus	JCM-15530	1
Lactobacillus similis	DSM-23365	1
Lactobacillus odoratitofui	DSM-19909	1
Lactobacillus collinoides	DSM-20515	1
Lactobacillus paracollinoides	DSM-15502	1
Lactobacillus malefermentans	DSM-5705	1
Lactobacillus parabuchneri	DSM-5707	1
Lactobacillus parabuchneri	DSM-15352	1
Lactobacillus buchneri	DSM-20057	1
Lactobacillus otakiensis	DSM-19908	1
Lactobacillus kefiri	DSM-20587	1
Lactobacillus parakefiri	DSM-10551	2
Lactobacillus sunkii	DSM-19904	2
Lactobacillus rapi	DSM-19907	1
Lactobacillus kisonensis	DSM-19906	1
Lactobacillus diolivorans	DSM-14421	1
Lactobacillus hilgardii	DSM-20176	1
Lactobacillus farraginis	DSM-18382	1
Lactobacillus parafarraginis	DSM-18390	2
Lactobacillus senioris	DSM-24302	1

Lactobacillus ozensis	DSM-23829	1
Lactobacillus kunkeei	DSM-12361	1
Lactobacillus florum	DSM-22689	1
Lactobacillus lindneri	DSM-20690	1
Lactobacillus sanfranciscensis	DSM-20451	1
Lactobacillus fructivorans	ATCC-27394	1
Lactobacillus homohiochii	DSM-20571	1
Lactobacillus fructivorans	DSM-20350	1
Lactobacillus fructivorans	DSM-20203	1
Pediococcus argentinicus	DSM-23026	0
Pediococcus claussenii	DSM-14800	0
Pediococcus pentosaceus	DSM-20336	0
Pediococcus stilesii	DSM-18001	0
Pediococcus lolii	DSM-19927	0
Pediococcus acidilactici	AS1-2696	0
Pediococcus ethanolidurans	DSM-22301	0
Pediococcus cellicola	DSM-17757	0
Pediococcus damnosus	DSM-20331	0
Pediococcus inopinatus	DSM-20285	0
Pediococcus parvulus	DSM-20332	0
Carnobacterium maltaromaticum	DSM-20730	0
Carnobacterium maltaromaticum	DSM-20342	0
Carnobacterium maltaromaticum	DSM-20722	0
Carnobacterium divergens	DSM-20623	1
Lactococcus lactis	LMG-7760	0
Atopobium minutum	DSM-20586	0
Olsenella uli	DSM-7084	2
Atopobium rimae	DSM-7090	1

Table S11. Genomic regions related to bacteriocin production identified in the 213 genomes.

Species Name	StrainID	Number of AOI identified in BAGEL	Bacteriocin homolog 1	CLASS	Confirmed in Artemis	Bacteriocin homolog 2	CLASS	Confirmed in Artemis	Bacteriocin homolog 3	CLASS	Confirmed in Artemis	Bacteriocin homolog 4	CLASS	Confirmed in Artemis	Bacteriocin homolog 5	CLASS	Confirmed in Artemis	Bacteriocin homolog 6	CLASS	Confirmed in Artemis	Bacteriocin homolog 7	CLASS	Confirmed in Artemis
Kandleria vitulina	DSM-20405	0																					
Lactobacillus kimchiensis	DSM-24716	1	Carnocin_CP52	Unmodified	Potential																		
Lactobacillus mindensis	DSM-14500	1	Carnocin_CP52	Unmodified	Potential																		
Lactobacillus nantensis	DSM-16982	2	Lactocin	Unmodified	Yes	Carnocin_CP52	Unmodified	Not a bacteriocin operon															
Lactobacillus crustorum	LMG-23699	0																					
Lactobacillus crustorum	JCM-15951	0																					
Lactobacillus futsaii	JCM-17355	1	Plantaricin like	Unmodified	Potential but across multiple contigs																		
Lactobacillus farciminis	DSM-20184	0																					
Lactobacillus alimentarius	DSM-20249	0																					
Lactobacillus paralimentarius	DSM-19674	0																					
Lactobacillus paralimentarius	DSM-13961	2	Sackacin /Enterocin	Unmodified	Yes	Carnocin	Unmodified	Not a bacteriocin operon															
Lactobacillus paralimentarius	DSM-13238	1	Carnocin_CP52	Unmodified	Potential																		
Lactobacillus tucceti	DSM-20183	0																					
Lactobacillus nodensis	DSM-19682	2	Carnocin_CP52	Unmodified	Not a bacteriocin operon	Gassericin A	Unmodified	Potential															
Lactobacillus versmoldensis	DSM-14857	0																					
Lactobacillus floricola	DSM-23037	0																					
Lactobacillus acidophilus	DSM-20079	3	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Lactacin f		Yes	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon												
Lactobacillus amylovorus	DSM-20531	7	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Potential	Helveticin J	Bacteriocin >10kd	Potential	Enterolysin A	Unmodified	Not a bacter	Unknown	Lanthipeptic	Potential	Gassericin K7	Unmodified	Not a bacter	Helveticin	Bacteriocin	Tranposase i
Lactobacillus amylovorus	DSM-16698	5	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Potential	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operc	Helveticin_J	Bacteriocin	Not a bacter	Helveticin_J	Bacteriocin	Not a bacteriocin operon						
Lactobacillus kitasatonis	DSM-16761	3	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin A	Bacteriocin >10kd	Not a bacteriocin operon												
Lactobacillus kefiranofaciens kefiranofaciens	DSM-5016	0																					
Lactobacillus kefiranofaciens kefirgranum	DSM-10550	3	Helveticin	Bacteriocin >10kd	Potential	Enterolysin	Bacteriocin >10kd	Potential but across multiple co	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon												
Lactobacillus ultunensis	DSM-16047	3	Helveticin J	Bacteriocin >10kd	Potential	Enterolysin	Bacteriocin >10kd	potential	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon												
Lactobacillus helveticus	LMG-22464	4	Enterolysin_A	Bacteriocin >10kd	Potential but across multiple contigs	Enterolysin_A	Bacteriocin >10kd	Potential but across multiple co	Helveticin J	Bacteriocin >10kd	Potential but across mu	Helveticin J	Bacteriocin	Not a bacteriocin operon									
Lactobacillus helveticus	CGMCC-1.1877	3	Enterolysin_A	Glyocin	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Potential	Helveticin_J	Bacteriocin >10kd	Potential												
Lactobacillus gallinarum	DSM-10532	4	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Potential but across multiple co	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operc	Helveticin_J	Bacteriocin	Not a bacteriocin operon									
Lactobacillus crispatus	DSM-20584	6	Helveticin	Bacteriocin >10kd	Not a bacteriocin operon	Thermophilin_A	Unmodified	Potential	Lactacin F	Unmodified	Yes	Enterolysin A	Bacteriocin	Not a bacter	Streptolysin	Bacteriocin	Potential but across multiple contigs						
Lactobacillus acetotolerans	DSM-20749	1	Enterolysin_A	Unmodified	Potential but across multiple contigs																		
Lactobacillus intestinalis	DSM-6629	6	Streptolysin	LAPs	Yes	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon	Planataracin	Unmodified	Potential	Helveticin_J	Bacteriocin	Not a bacter	Helveticin_J	Bacteriocin	Not a bacter	Enterolysin	Bacteriocin	Not a bacteriocin operon			
Lactobacillus hamsteri	DSM-5661	3	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin A/Lacto	Unmodified	Potential												
Lactobacillus amylolyticus	DSM-11664	3	Helveticin	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Potential but across multiple contigs												
Lactobacillus kalixensis	DSM-16043	6	Helveticin J	Bacteriocin >10kd	Potential	Enterolysin	Bacteriocin >10kd	Yes	Helveticin_J	Bacteriocin >10kd	Potential but across mu	Helveticin_J	Bacteriocin	Not a bacter	Lactococcin A	Unmodified	Not a bacter	Enterolysin	Bacteriocin	Not a bacteriocin operon			
Lactobacillus gigeriorum	DSM-23908	2	entetolysin_A	Bacteriocin >10kd	Potential but across multiple contigs	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon															
Lactobacillus pasteurii	DSM-23907	4	Enterolysin_A	Bacteriocin >10kd	Yes	Bovicin		Potenital	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operc	Helveticin J	Bacteriocin	Not a bacteriocin operon									
Lactobacillus delbrueckii jakobsenii	DSM-26046	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus delbrueckii lactis	DSM-20072	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus delbrueckii bulgaricus	DSM-20081	1	Enterolysin_A	Bacteriocin >10kd	Potential																		
Lactobacillus delbrueckii delbrueckii	DSM-20074	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus delbrueckii indicus	DSM-15996	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus equicursoris	DSM-19284	3	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Lactococcin	Unmodified	Potential	Helveticin_J	Bacteriocin >10kd	Potential												
Lactobacillus jensenii	DSM-20557	0																					
Lactobacillus psittaci	DSM-15354	0																					
Lactobacillus hominis	DSM-23910	1	Helveticin J	Bacteriocin >10kd	Yes																		
Lactobacillus taiwanensis	DSM-21401	3	Subtilin	Lanthipeptide_class_	Yes	Helveticin_J	Bacteriocin >10kd	Potenital	Lactacin F /Pediocir	Unmodified	Potential												
Lactobacillus johnsonii	ATCC-33200	3	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Helveticin_J	Bacteriocin >10kd	Not a bacteriocin operon	Pediocin	Unmodified	Yes												
Lactobacillus gasseri	ATCC-33323	0																					
Lactobacillus iners	DSM-13335	0																					
Lactobacillus amylotrophicus	DSM-20534	1	Enterolysin_A	Unmodified	Not a bacteriocin operon																		
Lactobacillus amylophilus	DSM-20533	1	Lactococcin_972	Unmodified	Yes																		
Lactobacillus dextrinicus	DSM-20335	0																					
Lactobacillus concavus	DSM-17758	0																					
Lactobacillus composti	DSM-18527	2	Enterolysin_A	Bacteriocin >10kd	Prophage	Pediocin	Bacteriocin >10kd	yes															
Lactobacillus harbinensis	DSM-16991	1	Plantaricin	Unmodified	Potential but across multiple contigs																		
Lactobacillus perolens	DSM-12744	0																					
Lactobacillus camelliae	DSM-22697	0																					
Lactobacillus nasuensis	JCM-17158	0																					
Lactobacillus manihotivorans	DSM-13343	0																					
Lactobacillus paracasei paracasei	DSM-5622	4	Unknown	Head_to_tail_cyclize	Not a bacteriocin operon	Gassericin A	Unmodified	Not a bacteriocin operon	Lactococcin A	Unmodified	Yes	Carnobacteric	Unmodified	Not a bacteriocin operon									
Lactobacillus casei	DSM-20011	3	Unknown	Head_to_tail_cyclize	Potential	Enterocin_X_chain_bet	Unmodified	Potential	Unknown	Unmodified	Not a bacteriocin operon												
Lactobacillus paracasei tolerans	DSM-20258	1	Unknown	Head_to_tail_cyclize	Potential																		
Lactobacillus zeae	DSM-20178	3																					
Lactobacillus rhamnosus	DSM-20021	2	Unknown	Head_to_tail_cyclize	Potential	Lactobin A/Cerein 7B	Unmodified	Yes															
Lactobacillus saniviri	DSM-24301	1	Enterolysin_A	Bacteriocin >10kd	Yes																		
Lactobacillus brantae	DSM-23927	0																					
Lactobacillus thailandensis	DSM-22698	0																					
Lactobacillus pantheris	DSM-15945	0																					
Lactobacillus sharpeae	DSM-20505	0																					
Lactobacillus selangorensis	DSM-13344	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus selangorensis	ATCC-BAA-66	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon																		
Lactobacillus graminis	DSM-20719	1	Lactococcin_972	Unmodified	Yes																		
Lactobacillus curvatus	DSM-20019	0																					
Lactobacillus sakei sakei	DSM-20017	0																					
Lactobacillus sakei carnosus	DSM-15831	0																					
Lactobacillus fuchuensis	DSM-14340	1	Sackacin	Unmodified	Not a bacteriocin operon																		
Lactobacillus rennini	DSM-20253	1	Pediocin Ach	Bacteriocin >10kd	Yes																		
Lactobacillus coryniformis torquens	DSM-20004	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Unknown	Head_to_tail_cycliz	Potenital															
Lactobacillus coryniformis coryniformis	DSM-20001	1	Unknown	Head_to_tail_cyclize	Potential																		
Lactobacillus bifermentans	DSM-20003	0																					
Lactobacillus ceti	DSM-22408	0																					
Lactobacillus saerimneri	DSM-16049	0																					
Lactobacillus animalis	DSM-20602	0																					
Lactobacillus murinus	DSM-20452	4	Hiracin_JM79	Unmodified	Not a bacteriocin operon	Plantaricin_S	Unmodified	Potential but across multiple co	Plantaricin_NC8	Unmodified	Potential but across mu	Enterolysin A	Unmodified	Not a bacteriocin operon									
Lactobacillus apodemi	DSM-16634	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Plantacarin	Unmodified	Yes															
Lactobacillus ruminis	DSM-20403	1	Pediocin-like	Unmodified	Potential but across multiple contigs																		

Lactobacillus agilis	DSM-20509	1	Plantacarin A	Unmodified	Yes							
Lactobacillus equi	DSM-15833	0										
Lactobacillus salivarius	DSM-20555	0										
Lactobacillus hayakitensis	DSM-18933	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Not a bacteriocin operon				
Lactobacillus pobuzihii	NBRC-103219	0										
Lactobacillus pobuzihii.Chen	KCTC-13174	0										
Lactobacillus acidipiscis	DSM-15836	3	Plantaricin_K /Carn	Unmodified	Potential but across multiple contigs	Pediocin/Colicin V	Unmodified	Potential but across multiple co	Pediocin/Colicin V	Unmodified	Potential but across multiple contigs	
Lactobacillus acidipiscis	DSM-15353	0										
Lactobacillus aviarius aviarius	DSM-20655	0										
Lactobacillus aviarius araffinosus	DSM-20653	0										
Lactobacillus sucicola	DSM-21376	0										
Lactobacillus aquaticus	DSM-21051	2	Mundticin_ATO6	Unmodified	Yes	Carnocin_CP52	Unmodified	Not a bacteriocin operon				
Lactobacillus uvarum	DSM-19971	0										
Lactobacillus capillatus	DSM-19910	0										
Lactobacillus cacaonum	DSM-21116	0										
Lactobacillus mali	DSM-20444	1	Carnocin_CP52	Unmodified	Yes							
Lactobacillus mali	ATCC-27304	0										
Lactobacillus hordei	DSM-19519	3	Carnocin_CP52	Unmodified	Potential	Plantacarin	Unmodified	yes	Coagulin	Bacteriocin >10kd	Not a bacteriocin operon	
Lactobacillus oeni	DSM-19972	0										
Lactobacillus satsumensis	DSM-16230	0										
Lactobacillus vini	DSM-20605	0										
Lactobacillus ghanensis	DSM-18630	0										
Lactobacillus nagelii	DSM-13675	0										
Lactobacillus algidus	DSM-15638	0										
Lactobacillus fabifermentans	DSM-21115	1	Carnocin_CP52	Unmodified	Yes							
Lactobacillus xiangfangensis	LMG-26013	2	Carnocin_CP52	Unmodified	Not a bacteriocin operon	Lactococcin_B	Unmodified	Potential but across multiple contigs				
Lactobacillus pentosus	DSM-20314	2	Plantacarin A	Unmodified	Yes	Pediocin	Unmodified	Not a bacteriocin operon				
Lactobacillus plantarum argentoratensis	DSM-16365	1	Enterocin		Potential but across multiple contigs							
Lactobacillus plantarum	DSM-13273	1	Plantaricin_K	Glyocin	Yes							
Lactobacillus plantarum plantarum	CGMCC-1.2437	1	Plantaricin_K	Glyocin	Yes							
Lactobacillus paraplantarum	DSM-10667	1	Plantaricin_K	Glyocin	Yes							
Lactobacillus siliginis	DSM-22696	0										
Lactobacillus rossiae	DSM-15814	2	Pediocin		Potential	Lactococcin	Unmodified	Potential but across multiple contigs				
Weissella viridescens	DSM-20410	0										
Weissella minor	DSM-20014	0										
Weissella halotolerans	DSM-20190	0										
Weissella confusa	DSM-20196	0										
Weissella kandleri	DSM-20593	0										
Oenococcus oeni	ATCC-BAA-1163	1	Unknown	Head_to_tail_cyclize	Potential							
Oenococcus kitaharae	DSM-17330	1	Streptolysin	LAPs	Yes							
Leuconostoc fallax	KCTC-3537	1	Lactobin A	Unmodified	Yes							
Leuconostoc pseudomesenteroides	4882	0										
Leuconostoc mesenteroides	ATCC-8293	0										
Leuconostoc mesenteroides cremoris	ATCC-19254	0										
Leuconostoc carnosum	JB16	0										
Leuconostoc argentinum	KCTC-3773	1	Lactococcin 972	Unmodified	Potential but across multiple contigs							
Leuconostoc citreum	KM20	1	Unknown	Head_to_tail_cyclize	Yes							
Leuconostoc gelidum	KCTC-3527	2	Enterocin	Unmodified	Potenital	Penocin_A	Unmodified	Yes				
Leuconostoc gasicomitatum	LMG-18811	1	Plantaricin NC8	Unmodified	Potential							
Leuconostoc kimchii	IMSNU-11154	1	Mesentericin B105	Unmodified	Yes							
Fructobacillus fructosus	DSM-20349	0										
Lactobacillus pontis	DSM-8475	0										
Lactobacillus panis	DSM-6035	1	Enterolysin_A	Bacteriocin >10kd	Potential							
Lactobacillus oris	DSM-4864	0										
Lactobacillus antri	DSM-16041	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Potential				
Lactobacillus reuteri	DSM-20016	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Bovicin		Not a bacteriocin operon				
Lactobacillus vaginalis	DSM-5837	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon				
Lactobacillus frumenti	DSM-13145	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon							
Lactobacillus fermentum	DSM-20055	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon							
Lactobacillus equigenerosi	DSM-18793	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Potential				
Lactobacillus gastricus	DSM-16045	1	Cytolysin A	Lanthipeptide_class_	Yes							
Lactobacillus ingluviei	DSM-15946	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Potential				
Lactobacillus ingluviei	DSM-14792	1	Enterolysin_A	Bacteriocin >10kd	Potential							
Lactobacillus secaliphilus	DSM-17896	0										
Lactobacillus coleohominis	DSM-14060	0										
Lactobacillus mucosae	DSM-13345	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Enterolysin	Bacteriocin >10kd	Not a bacteriocin operon				
Lactobacillus oligofermentans	DSM-15707	0										
Lactobacillus hokkaidonensis	DSM-26202	1	Leucocin A like	Unmodified	Yes							
Lactobacillus suebicus	DSM-5007	0										
Lactobacillus vaccinostercus	DSM-20634	0										
Lactobacillus parabrevis	LMG-11984	0										
Lactobacillus parabrevis	ATCC-53295	0										
Lactobacillus hammesii	DSM-16381	0										
Lactobacillus paucivorans	DSM-22467	0										
Lactobacillus senmaizukei	DSM-21775	0										
Lactobacillus brevis	DSM-20054	0										
Lactobacillus koreensis	JCM-16448	0										
Lactobacillus zymae	DSM-19395	0										
Lactobacillus acidifarinae	DSM-19394	0										
Lactobacillus namurensis	DSM-19117	0										
Lactobacillus spicheri	DSM-15429	0										
Lactobacillus kimchicus	JCM-15530	1	Pediocin	Unmodified	Yes							
Lactobacillus similis	DSM-23365	1	Unknown	Unknown	Potential but across multiple contigs							
Lactobacillus odoratitofui	DSM-19909	0										
Lactobacillus collinoides	DSM-20515	0										
Lactobacillus paracollinoides	DSM-15502	0										
Lactobacillus malefermentans	DSM-5705	0										
Lactobacillus parabuchneri	DSM-5707	1	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon							

Lactobacillus parabuchneri	DSM-15352	0												
Lactobacillus buchneri	DSM-20057	1	Unknown	Head_to_tail_cyclize	Potential									
Lactobacillus otakiensis	DSM-19908	1	Unknown	Head_to_tail_cyclize	Not a bacteriocin operon									
Lactobacillus kefiri	DSM-20587	0												
Lactobacillus parakefiri	DSM-10551	0	Unknown	Unmodified	Potential									
Lactobacillus sunkii	DSM-19904	0												
Lactobacillus rapi	DSM-19907	0												
Lactobacillus kisonensis	DSM-19906	0												
Lactobacillus diolivorans	DSM-14421	0												
Lactobacillus hilgardii	DSM-20176	0												
Lactobacillus farraginis	DSM-18382	1	Helveticin J	Bacteriocin >10kd	Not a bacteriocin operon									
Lactobacillus parafarraginis	DSM-18390	1	Helveticin J	Bacteriocin >10kd	Potenial									
Lactobacillus senioris	DSM-24302	0												
Lactobacillus ozensis	DSM-23829	0												
Lactobacillus kunkeei	DSM-12361	0												
Lactobacillus florum	DSM-22689	0												
Lactobacillus lindneri	DSM-20690	0												
Lactobacillus sanfranciscensis	DSM-20451	0												
Lactobacillus fructivorans	ATCC-27394	0												
Lactobacillus homohiochii	DSM-20571	0												
Lactobacillus fructivorans	DSM-20350	0												
Lactobacillus fructivorans	DSM-20203	0												
Pediococcus argentinicus	DSM-23026	0												
Pediococcus claussenii	DSM-14800	1	SalivaricinA	Lactipeptide	Potential									
Pediococcus pentosaceus	DSM-20336	1	Pediocin	Unmodified	Potential									
Pediococcus stilesii	DSM-18001	2	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon	Carnocin_CP52	Unmodified	Yes						
Pediococcus lolii	DSM-19927	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon									
Pediococcus acidilactici	AS1-2696	0												
Pediococcus ethanolidurans	DSM-22301	2	Pediocin like	Unmodified	Yes	Carnocin_CP52	Unmodified	Not a bacteriocin operon						
Pediococcus cellicola	DSM-17757	1	Carnocin_CP52	Unmodified	Potential									
Pediococcus damnosus	DSM-20331	2	Pediocin /Sacakcin	Unmodified	Yes	SalivaricinA	Lanthipeptide_class	Not a bacteriocin operon						
Pediococcus inopinatus	DSM-20285	1	Enterolysin_A	Bacteriocin >10kd	Not a bacteriocin operon									
Pediococcus parvulus	DSM-20332	1	Enterolysin_A	Bacteriocin >10kd	Prophage									
Carnobacterium maltaromaticum	DSM-20730	3	Piscicolin_126	Unmodified	Yes	Cytolysin_ClyLs	Lanthipeptide_class	Potential	Carnobacteriocin_B	Unmodified	Potential			
Carnobacterium maltaromaticum	DSM-20342	2	Carnocin_CP52	Unmodified	Potenital	Unknown	Sactipeptides	Not a bacteriocin operon						
Carnobacterium maltaromaticum	DSM-20722	4	Carnobacteriocin B:	Unmodified	Yes	Carnobacteriocin_BM1	Unmodified	Potential	Unknown	Sactipeptides	Not a bacteriocin operc	Cytolysin_Cl	Lanthipeptic	Potential
Carnobacterium divergens	DSM-20623	0												
Lactococcus lactis	LMG-7760	0	Plantacarin A	Unmodified	Not a bacteriocin operon									
Atopobium minutum	DSM-20586	0												
Olsenella uli	DSM-7084	0												
Atopobium rimae	DSM-7090	0												

Supplementary Notes

Supplementary Note 1

Lactobacillus relatedness to other genera

We also sequenced genomes of type strains from other genera associated with lactic acid bacteria (LAB) including *Carnobacterium* and *Lactococcus* and other genera previously misidentified as LAB, namely, *Atopobium*, *Kandleria* and *Olsenella*. *Carnobacterium* and *Lactococcus* form distinct branches and are classified in the families *Carnobacteriaceae* and *Streptococcaceae,* respectively, revealing closer genetic relatedness to genera other than *Lactobacillus*. *Carnobacterium* is closely related to *Melissococcus* and *Lactococcus* to *Streptococcus* (Fig. 1). This indicates that all currently known LAB descend from the same ancestor as proposed previously[1]. The un-relatedness of *Kandleria, Atopobium* and *Olsenella* and other LAB is confirmed, with a closer relationship of *Kandleria* with *Erysipelothrix* in the family *Erysipelotrichaceae* of the *Clostridium* subphylum cluster XVII of *Firmicutes. Atopobium* and *Olsenella* revealed the greatest genetic distance from the genus *Lactobacillus*; they belong to another phylum, *Actinobacteria,* which belongs to the high-GC, Gram-positive bacteria, the most closely related genus being *Coriobacterium*.

Supplementary Note 2

Validation of the core genome phylogeny

We validated the branching order of the 73 core gene tree using alternative subsets of gene datasets. We inferred a tree of the *Lactobacillus* genus complex, rooted on *Lactococcus lactis*, comprised of 117 genes, and another tree of the *Lactobacillus* genus complex, rooted on the *Carnobacteria*, comprised of 121 genes. We found there were only 4 minor branching alterations. The robustness of the 73 core gene tree was also assessed using a range of different models. Models tested were the WAG, LG, JTT, mtREV and Dayhoff models. The resulting branching order of the strains was consistent across all the models with the exception of mtREV. With this model the positions of *L. salivarius* DSM_20555, *L. hayakitensis* DSM_18933 and *L. acetotolerans* DSM_20749 underwent minor node exchanges within the same clade.

The impact of adding partial genes to the core gene dataset was assessed by inferring a phylogeny (Supplementary Fig. 19) from the 73 complete core genes with an additional 27 partial core genes (listed in Supplementary Table 8) added to the dataset. The inferred topology was highly congruent with the 73 complete core gene tree with only three very minor branching alterations. Specifically, *L. koreensis* moved one branch within the *L. brevis*/*L. collinoides* group. Minor displacements were also observed for *L. kefiranofaciens* DSM_5016 and DSM_10550 within the L*. delbrueckii* group and *L. salivarius* and *L. hayakitensis* within the *L. salivarius* group.

Supplementary Note 3

Relatedness of *Lactobacillus* species

At the species level, the combination of ANI value and core genome phylogeny could be proposed as the basis for optimal taxonomic classification[2], and its application could shed new light on several issues including species widely used in the food and probiotic industry. A long-debated case has been that of the *L. casei* group[3]. We propose that the species *L. casei* and *L. paracasei* should be combined into a single species, *L. casei*, because the pair-wise ANI values between the type strains of *L. casei* and two *L. paracasei* are 98~99%, larger than the "un-official" species' cut-off value of 95%[4], and these three strains clustered together as a monophyletic group (Fig. 2; Supplementary Fig. 20). The designation of *L. zeae* has been controversial and there are reports suggesting its classification into the species *L. casei*[3,5]. However, the ANI value between *L. casei* and *L. zeae* is only 77%~78%, which is similar to the value between *L. casei* and another well-defined species, *L. rhamnosus*. Therefore, our genomic analysis supports retaining *L. zeae* as a single species. Remarkably, ANI values seem not to be related to total DNA-DNA hybridization data reported previously[5] for similarities of *L. casei* type strain ATCC 393, which corresponds to the strain DSM 20011 included in this study.

The genus *Lactobacillus* was recently defined as 16 phylogroups (incl. *Pediococcus*), 4 couples (groups containing only two species) and 10 single species[6]. Such delineations are generally supported by the phylogenetic relationships constructed here based on the core proteins, but two modifications are suggested (Fig. 2; Supplementary Fig. 18). Firstly, *L. camelliae* could be included in the *L. manihotivorans* group and not in the *L. delbrueckii* group, supported by a mean TNI value of <15% to the group, which is far lower than that within the other *L. delbrueckii* group species (79.5%). Moreover, *L. amylotrophicus* and *L. amylophilus* should also be classified as a single species, *L. amylophilus*, as the two type strains have an ANI value of approximately 100%, although multilocus analysis and DNA-DNA hybridization values suggested their separation.

Supplementary Note 4

Phylogenomics of glycolysis and hexose fermentation

In species characterized by the presence of Pfk, the distribution of the pyruvate dehydrogenase operon (Pdh; composed of 4 genes) reflected carbohydrate metabolism since it is absent in 57% of obligately homofermentative species while it is present in facultatively heterofermentative members. Consistency was observed between Pdh distribution and phylogenetic groupings: 90% of the species of the *L. delbrueckii* group lack *pdh*, while members of groups like *L. salivarius*, *L. plantarum*, *L. casei* or *L. alimentarius* are characterized by presence of both *pfk* and *pdh,* although they belong to different phenotypic categorizations. Additionally, 80% of the species within the *L. delbrueckii* group lack

glucokinase, the first enzyme of the glycolytic pathway, except for the monophyletic subgroup formed by the *L. delbrueckii* and *L. equicursoris* species. For growth on glucose, a PTS transport system may obviate the need for this enzyme[7] and most of these species (79%) are classified as obligately homofermentative. Thus, comparative genomic analysis of glycolysis reveals that species inside the historically defined groups have a coherent genotypic background despite metabolic heterogeneity.

Supplementary Note 5

Carbohydrate active enzymes

Some GH families are present more uniformly across the dataset, indicating the importance of the biotransformations associated with these families. Included in this are enzymes involved in the hydrolysis of peptidoglycan (GH25 and GH73[8]), which play an important role in cell division, growth and preserving cell wall integrity. Bacterial autolysis can have a positive impact in the dairy fermentation process through the enhancement of cheese flavour upon the release of enzymes and amino acids[9]. Starch degradation enzymes are also present almost uniformly across the genome set. α-amylase enzymes are catalysts in the hydrolysis of the α-1,4 glycosidic linkages of starch with GH13 being the main GH family acting on substrates with α-glucoside linkages. α-glucan metabolism is important in the breakdown of resistant starch. Another universal family of GHs is GH65; this family is mainly composed of phosphorylases, including maltose and treholase phosphorylase, which are essential for the survival of lactobacilli in sugar-rich environments such as sourdough, where maltose is the predominant sugar[10].

Six clades display unusually high GH abundance, namely, *L. (par)alimentarius*, *L. perolens*, *L. plantarum*, *L. rapi*, *L. fructivorans* and *Carnobacterium* spp. The *Weissella* spp. and *L. fructivorans* clade show an unusually low GH gene count. The most abundant GH families are GH1, GH13, GH25 and GH73. GH73, an N-acetylmuramidase, is present in all *Lactobacillus* species except *L. equi*. The four *Carnobacterium* genomes harbor several GH families that are absent in all or most of the other genomes including GH18, GH24, GH84, GH85 and GH119. GH18 enzymes are chitinases, GH24 are lysozymes while GH119 enzymes are involved in chitin binding.

Some rare GTs emerge in the dataset. *P. lolii* and *P. parvulus* are the only strains that harbour genes for GT12, a ganglioside synthase. This activity is very interesting as it has been reported in only 3 bacterial species and 28 eukaryotes. Production of GT11 may be a strain-specific trait, because a single non-type strain of each of *L. johnsonii*, *L. amylovorus* and *L. paraplantarum* also appear to encode this activity (http://www.cazy.org/GT11_bacteria.html).

A number of GT families are present almost uniformly across the dataset. These include GT51, which is involved in peptidoglycan synthesis[8] and only *L. coleohominis* lacks

GT28, a galactosyltransferase involved in cell wall metabolism, suggesting alternative cell wall structure in this vaginal isolate. Other ubiquitous families include more broad-spectrum GTs that have been termed "polyspecific" because of their diverse functionality. Examples include GT4 and GT2, which encompass at least 12 functions including cellulose synthase, chitin synthase and mannosyltransferase[11].

The abundance of genes encoding carbohydrate transporters correlates strongly with GH abundance and less strongly with GT abundance (Supplementary Fig. 21). The clade distribution of carbohydrate transporter gene counts (Supplementary Fig. 22) mirrors this correlation and highlights the relative abundance of carbohydrate management machinery in the *L. alimentarius*, *L. casei*, *L. mali*, *L. plantarum* and *L. collinoides* clades, the pediococci and carnobacteria. Normalization for genome size (Supplementary Fig. 23) reduces the apparent overabundance of GT genes in some species, but not GH genes (Supplementary Fig. 23), and it is debatable if such normalization is biologically relevant.

Supplementary Note 6
Metabolic diversity of the lactobacilli

A general representation of the genome content of the lactobacilli as Clusters of Orthologous Groups (COGs; Supplementary Fig 15) reveals unexpected diversity in categories including Transcription, Cell wall biogenesis, Energy production, Co-enzyme transport, and Inorganic ion transport. The *L. plantarum*-related species from *L. fabifermentans* through *L. paraplantum* (in Fig. 2) are particularly endowed with genes for carbohydrate and amino acid transport and metabolism. Members of the *L. salivarius* clade have the highest number of genes involved in cell motility and secretion, as expected from previous studies in our group[12]. Forty nine (23%) of the genomes screened had none of the complete pathways for production of the 20 standard amino acids (AA), while no single genome harboured the genes to produce all 20 AAs. The highest number of pathways encoded by any one species was sixteen in *L. similis* (Supplementary Table 9). Some phylogenetic clades harbored genes for the production of one or two amino acids; these include *Weissella, L. brevis, Pediococcus* and *L. sakei*. The clade containing *L. collinoides* and *L. kimchicus* as well as the *L. coryniformis* clade (with the exception of *L. rennini*) are predicted to be prototrophic for at least 14 AAs. In contrast, ten of the twelve dairy isolates have genes for five or fewer complete AA pathways, reflecting their evolution to the AA-rich dairy environment.

Supplementary Note 7
Sortase-anchored proteins

The distribution of LPXTG proteins was not clearly correlated with the evolutionary relationship of lactobacilli. Rather, it seems that the decoration of the cell wall with anchored

proteins is a common feature, indicating that most lactobacilli can establish interactions with the environment. In our dataset of 213 genomes, only twelve bacterial strains do not have LPXTG proteins (Supplementary Table 5). In addition, most lactobacilli genomes harbor at least one gene encoding sortase, presumably the housekeeping sortase. The copy number of sortase genes could be associated to some extent with the presence of pilus gene clusters (PGC) (Fig. 4). These PGCs were identified in 51 strains in total, mostly belonging to six clades. The bacterial species with the most pilus gene clusters was *Lactobacillus sharpeae,* a sewage isolate. We also observed a large diversity of PGCs in terms of gene organization and numbers, suggesting they may have distinct gene order and functions (Supplementary Fig. 10). One of the most common PGC types consists of three pilin genes and one sortase gene, which is similar to the PGC originally reported in *L. rhamnosus* GG[13].

Supplementary Note 8

The *Lactobacillus* mobilome

The mobile component of the bacterial genome can expand coding capacity as in the case of plasmids and megaplasmids[14] or may be associated with genome decay as in the case of Insertion Sequence (IS) elements[15]. Some IS elements exhibit a very limited distribution across the 213 genomes, for example, IS1 which is restricted to only *L. ingluvei* and *L. equi* (Supplementary Fig. 16). IS91 is present in the genomes of only two species, both of dairy-product origin (*L. casei* and *L. paracasei* subsp. *tolerans*), and IS481 is found in only 3 strains (*L. paracollinoides*, *L. farraginis* and *P. inopinatus*) all associated with brewing. IS3, on the other hand, exhibits a much greater distribution being present in almost all groups, with the exception of *Weissella*/*Leuconostoc*, *L. fructivorans*, some *L. delbrueckii* strains and a few singletons. Some genomes apparently harbor no IS elements, perhaps indicating a more rigid architecture and a selective pressure against the acquisition of such elements. These species include *Atopobium minutum*, *L. fructivorans*, *L. florum*, *L. senoris*, *Weissella viridescens*, *L. cacaonum*, *L. apodemi*, *L. ceti* and *L. brantae*. Of the 18 IS families in this database, the largest number of families found is 13, in the *L. parabuchneri* genome. It is clear that IS elements have played a general role in shaping the diversification and evolution of the lactobacilli.

Phages were detected in the genomes of 195 of the 213 genomes (Supplementary Fig. 13), and only the genomes of *L. floricola*, *L. ingluviei*, *L. psittaci*, *L. sakei* subsp. *sakei*, *L. sanfranciscensis*, *P. cellicola*, *P. claussenii* and *W. halotolerans* lacked homologues of phage proteins. Proteins corresponding to holins and endolysins, and fibers/fiber assembly proteins were apparently under-represented; this is likely due to the fact that holins are not well annotated, whereas fiber proteins are very divergent, and not always present in the phage. It is likely that this analysis under-reports phage genes and that fine tuning of homology cut-offs for individual genes will identify more temperate phage. We identified plasmids in 41% of the

213 genomes analysed (Supplementary. Fig. 14) with numbers ranging from zero to six. 58% of the plasmid-related genes were of unknown function, emphasizing the need for functional genomics to elucidate this gene repertoire. Given the desirability of finding new plasmid vectors for genetically manipulating lactobacilli in the laboratory and for food-grade strain construction, these genomic data represent a valuable resource. Of the 87 strains predicted to have plasmids, 75 had identifiable replication genes. The other 12 may harbor new replication types that would be compatible with existing plasmid vectors.

Supplementary Note 9
Stress resistance

The broad range of niches that lactobacilli occupy is reflected in the multiple stress resistance mechanisms their genomes encode (Supplementary Fig. 24). Knowledge of the differential abundance of these systems can be exploited for identifying species and strains that can withstand production stress, storage stress or intestinal survival[16]. Superoxide dismutase and catalase show very limited distribution, as does the gene encoding glutamate decarboxylase involved in proton scavenging and acid resistance (Supplementary Fig. 24), the last of which is concentrated in *L. helveticus* and related species. Urease genes were present in only 8 species including 3 of the 4 *Carnobacteria* species. Bile salt hydrolases contribute to bile resistance in lactobacilli[17]; the pattern of their distribution in this genome data resource, viewed in conjunction with other stress resistance genes, allows rational identification of species likely to survive intestinal transit.

Protein acetyltransferases of prokaryotes like *Escherichia coli* confer resistance to heat and oxidative stress[18], and it was recently reported that homologs of the eukaryotic sirtuin protein acetyltransferases contribute to stress resistance in *Lactobacillus paracasei*[19]. Forty of the 213 genomes analyzed here lack homologs of any of SIRT1, SIR2L1, or Sir2α (Supplementary Table 10). The remaining 173 genomes encode at least one homolog, with a single species *L. ultunensis* harbouring 4 homologs, and 3 homologs being found in a large number of food or intestinal lactobacilli. Since pre-treatment of *Lactobacillus paracasei* strains with the sirtuin activator resveratrol (found in berries and red wine) alleviated growth inhibition by cholate[19], this presents the exciting prospect that some food ingredients might promote shelf-life of lactobacilli in functional food products, or that certain prebiotics might rationally promote survival and intestinal transit of lactobacilli administered in synbiotics.

Supplementary Note 10
Niche association and genome content

A search for associations between niche and genome content for the 213 genomes revealed moderate trends. The strongest trends were detected for species isolated from animal sources (n=56), which, as noted in the main text, also had the lowest number of

predicted genes and the smallest average genome size (Supplementary Fig. 25). This is evident at a functional level where numerous functional gene groups displayed the lowest abundances in the animal niche. These groups include genes for the transport and metabolism of carbohydrates (Supplementary Fig. 25, top panel), amino acids, lipids (Supplementary Fig. 25, top panel), co-enzymes, nucleotides and inorganic acids. Normalization for genome size brought the difference above the significance threshold (Supplementary Fig. 25, bottom panels; p-values of 0.11 and 0.107 for carbohydrate transport and lipid metabolism, respectively). Genes involved in transcription, cell wall/membrane biogenesis and secondary metabolites also have the lowest average abundance in the animal niche. Gene decay is commonly associated with host-adapted microbes and this is a possible explanation for the trend observed in genomes that have been isolated from animal sources.

Across the broader range of niches/sources, the analyses failed to detect niche-specific genomic associations in our study, but rather a general pattern of gene decay in species from the animal niche. This study focuses largely on the comparison of type strains within the genus *Lactobacillus* and associated species so it is likely that the diversity among species in the dataset was too great to reveal niche-specific traits. Studies that focus on multiple strains of the same species are more suited to discovering niche-specific genes.

Supplementary Note 11
Targeting competing microbes

Bacteriocins are small, ribosomally synthesized antimicrobial peptides that can be exploited as antimicrobial-producing cultures in fermented foods, or the bacteriocins themselves added, e.g. pediocin and carnobacteriocin used as biopreservatives[20]. Bacteriocins may also contribute to probiotic properties by limiting infection[21] or signalling to the innate immune system[22]. Of the 213 genomes analysed here, 107 (50.2%; Supplementary Table 11) harbored at least one Area of Interest (AOI) relating to bacteriocin production by screening against the BAGEL database[23]. Over half of these were larger proteins (>10 kDa) of the enterolysin/helveticin class, which are no longer considered classical bacteriocins[24]. However, their widespread distribution suggests a central function for these currently cryptic antimicrobials. Manual inspection confirmed that 38 AOIs had the contiguous gene structure expected for typical bacteriocin operons. No dominant bacteriocin type was identified, although 58% of intact AOIs fell under the Unmodified Bacteriocin (Class II) category. Many of these included homologs of well-known bacteriocin operons including plantaricin, sakacin, salivaricin, subtilin, leucocin, carnobacteriocin and lacticin F. Surprisingly, AOIs were identified in many species not previously associated with bacteriocin production. For example, genes encoding pediocins were annotated in *L. kimchicus* and *L. taiwanensis* was found to harbour the machinery to produce subtilin. Predicted bacteriocin loci were found across all

clades of the phylogenetic tree with the exception of *Weissella* and the *L. brevis/L.parabrevis* clade. Overall, the data do not support strict associations between species niche, the presence of a bacteriocin or type of bacteriocin. Moreover, the overall prevalence of bacteriocins was unexpectedly low compared to the literature describing this area, suggesting that many more *Lactobacillus* species can produce bacteriocins than those represented by their type strains examined in this study (for example *L. salivarius*[25]).

Supplementary References

1 Makarova, K. *et al.* Comparative genomics of the lactic acid bacteria. *Proc. Natl. Acad. Sci. U S A* **103**, 15611-15616, (2006).

2 Chan, J. Z., Halachev, M. R., Loman, N. J., Constantinidou, C. & Pallen, M. J. Defining bacterial species in the genomic era: insights from the genus Acinetobacter. *BMC Microbiol* **12**, 302, (2012).

3 Judicial Commission of the International Committee on Systematics of Bacteria. The type strain of Lactobacillus casei is ATCC 393, ATCC 334 cannot serve as the type because it represents a different taxon, the name Lactobacillus paracasei and its subspecies names are not rejected and the revival of the name 'Lactobacillus zeae' contravenes Rules 51b (1) and (2) of the International Code of Nomenclature of Bacteria. Opinion 82. *Int J Syst Evol Microbiol* **58**, 1764-1765, (2008).

4 Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81-91, (2007).

5 Dicks, L. M., Du Plessis, E. M., Dellaglio, F. & Lauer, E. Reclassification of Lactobacillus casei subsp. casei ATCC 393 and Lactobacillus rhamnosus ATCC 15820 as Lactobacillus zeae nom. rev., designation of ATCC 334 as the neotype of L. casei subsp. casei, and rejection of the name Lactobacillus paracasei. *Int J Syst Bacteriol* **46**, 337-340, (1996).

6 Salvetti, E., Torriani, S. & Felis, G. E. The genus *Lactobacillus*: a taxonomic update. *Probiotics Antimic. Proteins* **4**, 217-226, (2012).

7 Grossiord, B., Vaughan, E. E., Luesink, E. & de Vos, W. M. Genetics of galactose utilisation via the Leloir pathway in lactic acid bacteria. *Lait* **78**, 77-84, (1998).

8 Cayrou, C., Henrissat, B., Gouret, P., Pontarotti, P. & Drancourt, M. Peptidoglycan: a post-genomic analysis. *BMC Microbiol.* **12**, 294, (2012).

9 Huard, C. *et al.* Characterization of AcmB, an N-acetylglucosaminidase autolysin from *Lactococcus lactis*. *Microbiology* **149**, 695-705, (2003).

10 Ehrmann, M. A. & Vogel, R. F. Maltose metabolism of *Lactobacillus sanfranciscensis:* cloning and heterologous expression of the key enzymes, maltose phosphorylase and phosphoglucomutase. *FEMS Microbiol. Lett.* **169**, 81-86, (1998).

11 Breton, C., Snajdrova, L., Jeanneau, C., Koca, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R-37R, (2006).

12 Neville, B. A. *et al.* Characterization of Pro-Inflammatory Flagellin Proteins Produced by Lactobacillus ruminis and Related Motile Lactobacilli. *PLoS One* **7**, e40592, (2012).

13 Kankainen, M. *et al.* Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. *Proc Natl Acad Sci U S A* **106**, 17193-17198, (2009).

14 Li, Y. *et al.* Distribution of megaplasmids in *Lactobacillus salivarius* and other lactobacilli. *J. Bacteriol.* **189**, 6128-6139, (2007).

15 Callanan, M. *et al.* Genome sequence of *Lactobacillus helveticus*, an organism distinguished by selective gene loss and insertion sequence element expansion. *J. Bacteriol.* **190**, 727-735, (2008).

16 Zhang, Y. & Li, Y. Engineering the antioxidative properties of lactic acid bacteria for improving its robustness. *Curr. Opin. Biotechnol.* **24**, 142-147, (2013).

17 Fang, F. *et al.* Allelic variation of bile salt hydrolase genes in *Lactobacillus salivarius* does not determine bile resistance levels. *J. Bact.* **191**, 5743-5757, (2009).

18 Ma, Q. & Wood, T. K. Protein acetylation in prokaryotes increases stress resistance. *Biochem. Biophys. Res. Commun.* **410**, 846-851, (2011).

19 Atarashi, H. *et al.* in *LAB11 2014, poster B014.* (Egmon aan Zee, Netherlands, 2014).

20 Cotter, P. D., Hill, C. & Ross, R. P. Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.* **3**, 777-788, (2005).
21 Corr, S. C. *et al.* Bacteriocin production as a mechanism for the antiinfective activity of *Lactobacillus salivarius* UCC118. *Proc. Natl. Acad. Sci. U S A* **104**, 7617-7621, (2007).
22 Meijerink, M. *et al.* Identification of genetic loci in Lactobacillus plantarum that modulate the immune response of dendritic cells using comparative genome hybridization. *PLoS One* **5**, e10632, (2010).
23 de Jong, A., van Heel, A. J., Kok, J. & Kuipers, O. P. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.* **38**, W647-651, (2010).
24 Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108-160, (2013).
25 O'Shea, E. F. *et al.* Production of multiple bacteriocins from a single locus by gastrointestinal strains of *Lactobacillus salivarius*. *J. Bacteriol.* **193**, 6973-6982, (2011).

Appendix to Chapter III

Published as Supplementary:

Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal.

Harris HMB, Bourin MJB, Claesson MJ, O'Toole PW.

Microb Genom. 2017 Jun 13; 3(8): e000115.

doi: 10.1099/mgen.0.000115.

eCollection 2017 Aug.

PMID: 29026656

SUPPLEMENTARY FIGURE LEGENDS

Fig. S1: A phylogenetic tree of the *parB* gene showing similarity in tree topology compared to the core-gene tree in Fig. 2. Branch lengths (solid black lines) represent evolutionary divergence and strain labels are lined up for ease of comparison (dashed lines). Bootstrap values are included to show robustness of tree topology. The tree is rooted on *L. hayakitensis* DSM18933 and this branch is artificially reduced to provide a clearer visualisation of the other branch lengths relative to each other. The scale bar shows average number of amino acid substitutions per site.

Fig. S2: A synteny-based representation of EPS 2 shows considerable variation in the central genes. *L. salivarius* strains possessing EPS 2 have been separated into 6 groups depending on their synteny with UCC118 EPS 2. The order and colour of the genes above the groups agree with the order and colour of the genes in Fig. 8 (which includes a colour legend providing functional descriptions for each gene).

Fig. S3: Clustering of pair-wise percentage of conserved proteins (POCP) scores shows inconsistent clustering of sub-clades compared to the core-gene tree in Fig. 2. The colour key (top-left) shows a gradation of colour from red to orange to yellow to white representing increasing genome-genome similarity. Euclidean distance and complete linkage clustering were used to cluster rows and columns. *L. hayakitensis* DSM18933 is excluded.

Fig. S4: Gene counts for protease families (MEROPS database) have very different distributions across the strains. Predicted families of protease inhibitors have also been included. The colour key (top-left) shows a gradation of colour from blue to dark blue to purple, representing increasing protease gene counts. Grey represents a gene count of zero. Protease family annotations are listed in Table S5. The order of strains from left to right (columns) reflects the order of the core-gene tree from top to bottom in Fig. 2.

Fig. S5: Gene counts of COG categories for predicted prophages show that some strains lack intact prophages. The colour key (top-left) shows a gradation of colour from blue to dark blue to purple, representing increasing COG gene counts. Grey represents a gene count of zero. The order of strains from left to right (columns) reflects the order of the core-gene tree from top to bottom in Fig. 2. The barplot shows the number of predicted prophages based on the predictions of Virsorter (with the least confident category excluded).

Fig. S6: Some strains do not have any intact predicted prophages. The barplot shows the number of predicted prophages based on the predictions of Virsorter (with the least confident category excluded).

Fig. S7: A phylogenetic tree generated from the cas 1 gene from type-II CRISPR-cas systems shows agreement with some sub-clades from the core-gene tree in Fig. 2. Branch lengths (solid black lines) represent evolutionary divergence and strain labels have been lined up for ease of comparison (dashed lines). Bootstrap values are included to show robustness of tree topology. The scale bar shows average number of amino acid substitutions per site.

Fig. S8: Gene counts for 14 IS families show different distributions over both strains and replicons. The colour key (top-left) shows a gradation of colour from blue to dark blue to purple as the gene count increases. Grey represents a gene count of zero. IS genes have been separated into chromosomal, megaplasmid and plasmid genes. For each replicon group, the order of strains from left to right (columns) reflects the order of the core-gene tree from top to bottom in Fig. 2.

Fig. S9: Gene counts for signal peptides and transmembrane proteins. The order of strains from left to right (bars) reflects the order of the core-gene tree from top to bottom in Fig. 2.

Fig. S10: Pair-wise BLAST scores of each predicted bile salt hydrolase gene (Bsh) against the others (blastp) reveals three main clusters. The colour key (top-left) shows a gradation of colour from red to orange to yellow as the bit score increases. Genes have been separated into chromosomal (green) and megaplasmid (blue) groups. There were no BSHs predicted on smaller plasmids. Euclidean distance and complete linkage clustering were used to cluster rows and columns. Row labels have been excluded to avoid making strain names appear squashed, but they have a top-down order mirroring the column labels from left to right.

Fig. S11: Gene counts for antibiotic resistance genes (AR) and virulence factors (VF) show considerable intra-specific variation. The colour legend separates ARs from VFs. The order of strains from left to right (bars) reflects the order of the core-gene tree from top to bottom in Fig. 2.

SUPPLEMENTARY TABLE LEGENDS

Table S1: Genome summary statistics. The number of predicted genes for both complete and draft genomes is based on the methods used in our study (see Methods). The number of small plasmids is an estimate based on the homology of contigs to one or several complete plasmids downloaded from NCBI. It should be noted that a plasmid count based solely on homology is unreliable and our study focuses on the functional capacity of plasmids rather than on their copy number. Details concerning linear megaplasmids and multiple circular megaplasmids are covered in Results and Discussion.

Table S2: BLAST results for *repA* (LSL_1739), *repE* (LSL_1740) and *parA* (LSL_1741) show consistent hits to contigs assigned to megaplasmids. Results were filtered at 60% identity and 50% alignment length of the query gene.

Table S3: The presence of putative bacteriocins shows both chromosomal and extrachromosomal variation. For each bacteriocin, its class and predicted replicon are also listed. Column 2 shows the Area of Interest (AOI; Bagel3), which describes a genomic region with one or more bacteriocins surrounded by one to several marker genes for that bacteriocin. Multiple bacteriocins from one strain within the same AOI are in close proximity.

Table S4: Results of Kraken show a lack of obvious contamination in *L. salivarius* genomes. Columns from left to right show strain name, taxon classification, percentage of nucleotides assigned to each taxon, number of nucleotides assigned to each taxon and contig count per taxon. Note that if a contig is assigned to a particular taxon, all nucleotides within that contig are counted as being part of the same taxon.

Table S5: Sensitivity and specificity values for assigning artificial contigs to the NCBI plasmid database. For each genome, there are 5 rows and 2 columns – the rows represent all replicons shared among the 4 genomes, the first column shows correct replicon assignment and the second shows incorrect replicon assignment. Technically, the first column is sensitivity and the second is the false negative rate (1 – specificity). Values represent percentage of assigned nucleotides and NA mean that a specific replicon is absent from that genome. Megaplasmid 1 represents the *repA*-type megaplasmid of *L. salivarius*.

Table S6: Annotations for the protease families from Fig. S4. This information has been reproduced from the MEROPS website (http://www.merops.sanger.ac.uk).

Table S7: Predicted CRISPRs are all harboured on the chromosome and vary in type across the strains. Columns 1-10 show genome, CRISPR type, CRISPR sub-type, repeat number, repeat length, repeat sequence, average spacer length and marker cas gene (cas 1 for all types; cas 9 for type-II; cas 10 for type-III). The absence of CRISPRs for a genome is denoted by NA. Undefined CRISPRs have predicted repeat and spacer sequences, but no predicted cas genes.

Table S8: BLAST results for antibiotic resistance genes against the CARD database. Results were filtered at 40% identity and 50% alignment length of the query gene.

Table S9: BLAST results for virulence factors against the VFDB database. Results were filtered at 70% identity and 90% alignment length of the query gene.

MICROBIAL GENOMICS

Research paper template

SUPPLEMENTARY TEXT

METHODS

Quality assessment of genomes

Three additional quality control steps were performed on all genomes. First, the genomes were BLASTed (blastn v2.2.26+) [1] against a filtered version of the RDP database (v11.1) [2] that included only complete or near-complete 16S rRNA genes (>1400 bp) annotated to species level. This was carried out to confirm that the top hit for each assembly was an *L. salivarius* sequence in the RDP database and also to make sure that no other good hit (>1000 bp; identity >= 97%) was found, which would indicate possible contamination. Second, 39 universal marker genes [3] were BLASTed (tblastn) against the contigs of each genome. The reasoning here was that all 39 marker genes should be fully assembled in a high-quality genome (which they were). Third, the contigs of each genome were assessed using Kraken (v0.10.6) [4] as a further test for possible contamination. The results of all 43 Kraken runs are shown in Table S4.

Assigning contigs to replicons

A common problem when analysing genes of interest in draft genomes is being able to tell whether a particular gene is present on the chromosome or on a plasmid. Since all of a contig must either be part of the chromosome or part of a plasmid, once a contig has been assigned to a replicon the genes present on the contig can also be assigned to the same replicon. We describe here the method used in this study to assign each contig to its most likely replicon.

A database of 92 plasmids (10 megaplasmids and 82 plasmids) from 16 *Lactobacillus* species was generated from complete genomes available on NCBI. Each draft genome was BLASTed (blastn v2.2.26+) against this database and the results were filtered in order to assign each contig to a plasmid or, failing plasmid assignment, to the chromosome. Megaplasmids (>10 kb; both circular and linear) and smaller plasmids (<10 kb) were included in the database so four categories of replicon were possible (including the chromosome).

To justify BLAST thresholds for assigning contigs to replicon categories, 4 complete genomes (UCC118, CECT5713, Ren and NIAS840) were broken up into contigs using a randomly generated chi-squared distribution with Degrees of Freedom equal to 1. This distribution was chosen because its median Spearman correlation with the distribution of lengths of contigs for each draft genome in our study was 0.96 (Q1 = 0.9; Q3 = 0.98; n = 38), ensuring that the artificial draft genomes would resemble the draft genomes in our dataset in terms of contig length distributions. The values of this distribution were then converted to proportions and randomly permuted in order to avoid a bias between contig length and genome region. Each genome was divided up based on the order of the randomly permuted proportions where each proportion is a fraction of the total number of base pairs in the genome. For each of the 4 complete genomes, each replicon was broken up into 50 contigs and contigs less than 200 bp were excluded. The FASTA header for each contig was labelled with the

replicon from which it was taken so that the specificity and sensitivity of the contig assignment method could be tested. The R (v3.2.3) code uploaded to figshare (Data Bibliography of main text; data file 5) shows the steps for generating 50 draft contigs from the complete chromosome sequence of UCC118.

The 4 artificial draft genomes were then BLASTed (blastn) against the database of 92 plasmids. This was done for each genome separately so that the complete plasmids from each genome being BLASTed could be removed from the database beforehand. This ensured that draft plasmid contigs were not just aligning to the complete version of their own plasmids. An unfiltered evaluation of the BLAST results showed that the highest % alignment length of a chromosomal contig against the plasmid database was 23.7% (490/2,070). An alignment length of 25% against the plasmid database was therefore chosen as the cut-off for assigning contigs to plasmids. BLAST hits between two sequences can have multiple high-scoring pairs (HSPs) so the sum of the non-overlapping length of all HSPs between each contig and reference plasmid was calculated. The reference plasmid sequence with the highest % alignment to the contig was chosen and all alignments of less than 25% were excluded. Depending on their top hit, these remaining contigs were assigned to one of three categories: plasmid, circular megaplasmid or linear megaplasmid. It should be noted that small contigs representing transposases or other small repetitive regions may be present on both the chromosome and the plasmid(s) so the assignment of these contigs is less reliable. The sensitivity and specificity of the BLAST results for the 4 artificial draft genomes against the plasmid database are shown in Table S5. Code for calculating the sum of the non-overlapping length of HSPs between each contig and each reference plasmid has been uploaded to figshare (Data Bibliography of main text; data file 6).

As an additional quality check, three genes identified as being specific to the *L. salivarius* circular megaplasmid(5) - *repA* (LSL_1739), *repE* (LSL_1740) and *parA* (LSL_1741) – were BLASTed (tblastn) against the contigs for each genome assembly to see if all top hits were to predicted megaplasmid contigs. Results are in Table S2.

Specific functional groups

COG categories for genes were predicted by BLASTing (blastp) amino acid sequences against a COG database (ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data) with thresholds of 40% identity, 50% alignment length of the query gene and a BLAST bit score of 60. Any gene match that fell below these thresholds was added to the COG category ‘unknown function’. For each genome, genes were assigned to their respective replicons.

Peptidases were predicted by BLASTing (blastp) amino acid sequences against full sequences from the MEROPS database (https://merops.sanger.ac.uk). BLAST thresholds used were 40% identity, 50% alignment length of the query gene and a BLAST bit score of 60.

Sortase genes were predicted using hmmscan from the HMMER3 (v3.1b1) [6] toolkit with the following downloaded sortase family HMM profiles: http://nihserver.mbi.ucla.edu/Sortase/sortase_family_classification.hmm. A cut-off score of >30 was chosen to balance false positive and false negative predictions based on comparisons with the non-redundant NCBI and KEGG annotations. Genes with an LPXTG motif were predicted using hmmscan with a TIGRFAM [7] HMM profile (TIGR01167) and an e-value cut-off of 1e-05. The LOCP [8] webserver was used to locate putative pilus operons using default parameters. All three methods used amino acid sequences as input.

Glycosyl hydrolases and glycosyl transferases were predicted using hmmscan with DBcan [9] HMM profiles (http://csbl.bmb.uga.edu/dbCAN/). For each genome, GH and GT genes were assigned

to their respective replicons. A cut-off score of >30 was chosen to balance false positive and false negative predictions based on comparisons with non-redundant NCBI and KEGG annotation.

The Bagel3 [10] webserver was used to predict genetic loci for bacteriocin production and surrounding areas of interest (AOIs) using marker genes. For each genome, AOIs were sorted into their respective replicons.

CRISPRs were predicted using MinCED (v0.2.0), which was downloaded from the following link: https://github.com/ctSkennerton/minced. To predict cas genes associated with each CRISPR, hmmscan was used with cas-specific HMMs from TIGRFAM. CRISPRs that had no associated cas genes were labelled as 'undefined'. The same method used to build the core-gene phylogenetic tree was also used to build a tree from the amino acid sequences of the cas 1 (type-II and type-III) gene.

Genes involved in exopolysaccharide biosynthesis were predicted using the two EPS clusters of UCC118 as references. This was the only functional group that relied on a reference genome and it was used in order to give an overview of EPS genetic diversity in *L. salivarius* since a much larger, more detailed study is being conducted on the intra-specific diversity and functionality of EPS clusters in *L. salivarius* (Bourin *et al*; in preparation). Amino acid sequences of the UCC118 EPS genes were BLASTed (tblastn) against contigs with thresholds of 40% identity, 50% alignment length of the query gene and a BLAST bit score of 60. BLAST hits were then manually curated, taking note of UCC118 EPS genes present in multiple copies (in the case of transposases) and genes that passed the thresholds but were located in very different regions of the genome than the other predicted EPS genes.

Signal peptides were predicted using SignalP (v4.1) [11] with default parameters for gram-positive bacteria. Transmembrane domains were predicted using TMHMM (v2.0) [12] and all predictions with more than 10 expected amino acids in transmembrane helices in the first 60 amino acids were excluded from the results due to their likelihood of being signal peptides (see Instructions at http://www.cbs.dtu.dk/services/TMHMM/).

Antibiotic resistance (AR) genes were predicted using an AR reference gene set from the Comprehensive Antibiotic Resistance Database (CARD; v1.09) [13]. Within the CARD database, the FASTA file denoted 'protein homolog model' was filtered to include only complete genes and then genes were translated from nucleotide to amino acid sequences. Amino acid sequences for each genome were BLASTed against this database and filtered at 40% identity, 50% alignment length of the query gene and a BLAST score of 60.

Potential virulence factors (VF) were predicted using a version of the virulence factor database (VFDB) [14], which was downloaded from the following link: http://www.mgc.ac.cn/VFs/Down/VFDB_setA_pro.fas.gz. This database is the core dataset and contains virulence factor genes that have been experimentally verified only - the full database was not used in order to minimise the number of false positive gene predictions. Amino acid sequences for each genome were BLASTed against the database and filtered at 70% identity and 90% alignment length of the query gene. More stringent cut-off values were used for virulence factors compared with antibiotic resistance genes because using BLAST to identify homologous genes based on the VF database is known to produce false positives at lower cut-off values.

Prophages were predicted using VIRSorter (v1.0.2) [15] where predicted regions with the lowest confidence (category 3; 'not so sure') for both complete phage contigs and prophages were excluded. Predicted phage genes for the remaining categories were assigned to COG categories (the same COG database used for the general COG analysis) using blastp with thresholds of 40% identity, 50% alignment length of the query gene and a BLAST bit score of 60. Any gene that fell below these thresholds was added to the COG category 'unknown function'.

Transposases were predicted using hmmscan with TnpPred [16] HMM profiles downloaded from the following link: http://www.mobilomics.cl/. An e-value cut-off of 1e-05 was used. For each genome, predicted transposases were assigned to their respective replicons.

Bile salt hydrolase genes were predicted using a subset of the KEGG database where the EC number 3.5.1.24 was used to select bile salt hydrolase genes. Amino acid sequences were BLASTed (blastp) against this database with thresholds of 40% identity, 50% alignment length of the query (and reference) gene and a BLAST bit score of 60. These genes were then BLASTed against each other to give a pairwise BLAST score for each pair of bile salt hydrolase genes.

All statistics and data visualisation were carried out in R (v3.2.3) [17]. R packages used during this study were MADE4 [18] and SeqinR [19].

RESULTS AND DISCUSSION

The core-gene phylogenetic tree of *L. salivarius* has similar sub-clade topology to POCP clusters, but overall tree topology is dissimilar

Percentage of Conserved Proteins (POCP) [20] calculates a similarity score based on percentage of genes in common between all the amino acid sequences in two genomes. POCP was designed as a method to identify whether a particular species belongs within a genus. We were not interested in applying this threshold since all strains obviously fall within a single genus; instead, the goal was to assess the congruency of a core-gene phylogeny with a method that clustered the strains based on the presence and absence of genes. Fig. S3 shows a heatmap of POCP values, where clustering of strains is in reasonable agreement with the core-gene phylogeny of Fig. 2 in terms of sub-clades. Several strains cluster apart from their core-gene sub-clades including CECT5713 and CCUG38008. A greater difference between POCP and the core-gene tree versus ANI and the core-gene tree is expected because POCP value calculations ignore homologous regions, using similarity based on gene presence and absence distributions to cluster strains. This is a rough approximation of the combined effect of gene decay and HGT since a gene that is present in one strain and absent in another has either acquired a deleterious mutation or else has been horizontally transferred by one of several mechanisms. The reason why many of the sub-clades in Fig. S3 agree with the core-gene phylogeny is that the probability of gene decay or HGT events having occurred after two strains start to diverge from a common ancestor increases with time. Adaptation to different niches and differing selection pressures then start to disrupt the correlation between core-gene phylogeny and clustering of shared/unshared genes [21]. We found no general association of clusters from any tree generated in this study with the isolation sources of the strains (Table S1), but members of several small clusters were all isolated from the same source. This overall lack of niche-strain association may be due to the transient appearance of *L. salivarius* in niches associated with the gastro-intestinal tract (food, opportunistic infection of body sites, etc.) and it would be a mistake to assume that every strain has acquired niche-specific adaptations to its source of primary isolation.

Protease genes show no variation or considerable variation depending on MEROPS protease family

Proteases are a large group of proteins, divided into many families that are involved in the hydrolysis of peptides. Fig. S4 shows 53 protease families that display variation across the 43

genomes in this dataset. Genes for eighteen additional protease families were predicted in *L. salivarius*, but these families showed no variation across the strains, with 17 represented by a single gene per strain (A01A, C108, C14B, C19, C46, I04, I87, M02, M10A, M13, M15D, M20B, M20D, M24A, S09A, S09B, T05, T06) and one, a cysteine protease (C19) described as 'ubiquitin-specific' by the MEROPS database (Table S6), represented by two genes per strain. It can be speculated that these 18 families are subjected to purifying selection since the remaining 53 protease families vary both in gene count and in presence and absence across the strains.

Out of the 53 protease families that vary in their distributions, 33 are present in all 43 genomes, but have variable gene counts; genes for thirty of these are found on the chromosome only while the remaining three are present on multiple replicons. The gene count per protease family ranges from 0 to 24 where some families are present in all but a single genome and other families are present in one only (usually DSM18933 – the strain of *L. hayakitensis* used in this study). The protease family with the most genes in *L. salivarius* (4-24) is M23B, which is annotated as a lysostaphin in the MEROPS database (Table S6), an antibacterial enzyme that degrades peptidoglycan in the cell walls of certain bacteria, staphylococci in particular.

There are a number of protease families and protease inhibitors that are rare in the dataset of *L. salivarius* annotations, with representatives belonging to one or several genomes only. JCM1046 has gene products in two families that the other strains do not have – I75 and S26B – both relevant genes predicted to reside on the chromosome. The gene encoding I75 is on a small contig of 964 bp that has a 99% match over its full length to a phage from *E. coli*, suggesting recent acquisition of this sequence as a prophage. The only other predicted protease inhibitor, I63, is an inhibitor of pappalysin-1 and it is present in all *L. salivarius* genomes but absent from *L. hayakitensis* DSM18933. S26B is a signal peptidase that cleaves signal peptides from a secreted protein as it is being translated. DSM18933 has 2 protease families that are not present in *L. salivarius*, which suggests that they were either horizontally acquired by *L. hayakitensis* after the split from its common ancestor with *L. salivarius* or else that *L. salivarius* subsequently lost these families through gene decay, whether through genetic drift or active selection pressure. These two families are M42 and M60, a glutamyl aminopeptidase and an enhancin, respectively.

Sun *et al* conducted a genus-wide, comparative genomic study of lactobacilli and found considerable variation in cell-envelope proteases [22]. Our study shows that a more general overview of protease families reflects the high levels of variation seen in *Lactobacillus*, at the species level, in *L. salivarius*.

Prophages, CRISPRs and insertion sequences are widely distributed across *L. salivarius* but no obvious association exists between them

Two agents of HGT that affect both the bacterial chromosome and extrachromosomal replicons are bacteriophages and insertion sequences (consisting primarily of a transposase gene). Bacteriophages are ubiquitous among bacterial communities and phage-host dynamics has been shown to stabilise diversity within a community [23] as well as to drive the arms race between the evolution of bacterial defences (often in the form of CRISPR-cas systems) and the counter-evolution of phage structures that neutralise those defences [24]. Insertion sequences (IS) have been implicated in the horizontal transfer of a wide range of functions and are noted for their role in conferring niche-specific advantages to bacteria, allowing the persistence of strains or species in new environments that were previously uninhabitable [25].

Fig. S5 shows a heatmap of predicted prophage genes (COGs) and Fig. S6 shows a barplot of prophage counts. Nine strains (2 sub-clades of 4 strains each and JCM1230) lack predicted

prophages; it is unlikely that these 9 strains have no history of interacting with bacteriophages - instead, VirSorter has failed to predict relatively intact prophages in the genomes of these strains. Canchaya *et al* summarise the relationship between bacteria and prophages by writing that prophages are lost from bacterial genomes as easily as they are acquired [26]. There is no clear association of the two sub-clades with a single niche, although the human oral cavity is the isolation source of 5 of these strains and the other 4 were isolated from the mammalian intestine. It is tempting to suggest that the oral environment selects against the persistence of prophages; however, Edlund *et al* describe the oral cavity as the perfect portal for viruses to access the oral microbial community [27] and previous studies have shown that it is host to a diverse community of phages [28, 29].

The COG category in Fig. S5 with by far the most genes is 'Function unknown' (S) with a mean average gene count of 61.3 compared with the second highest - 'General function prediction only' - of 3. The size of these categories emphasises the limits of current knowledge regarding bacteriophage gene function. There is a correlation between number of predicted prophages and number of prophage genes (Spearman; rho = 0.78; $p < 0.001$), which is largely expected and highlights the size constraints on phages that infect *L. salivarius* since number of prophages, not phage type, approximately accounts for number of prophage genes. Some of the COG categories that are least abundant in predicted prophages are those involved in cell-specific functions such as cell motility (N) and secretion (U). The distribution of the remaining COG categories across the strains is indicative of the dynamic nature between bacteria and their phages, with considerable intra-species variation suggesting that the prophage complement of the ancestor of *L. salivarius* does not resemble any of the currently extant strains since their repertoire of prophages is so distinct.

Table S7 describes the distribution of CRISPRs across the 43 genomes as well as their associated cas genes. All CRISPRs are located on the chromosome, highlighting their role in protecting against extrachromosomal sequences. Almost all strains in this dataset have either the type-II or type-III CRISPR-cas system (or both), identified by the cas 9 or cas 10 gene, respectively, and 6 strains have no identified CRISPRs. The presence of either type-II or type-III CRISPR-cas systems show some clustering on the core-gene tree in Fig. 2: the DSM20555^T sub-clade consisting of 4 strains all have the type-III system only while the CECT5713 (6 strains) and UCC118 (4 strains) sub-clades have the type-II system only; the AH43348 sub-clade (6 strains), in contrast, has both type-II and type-III systems. The partial clustering of CRISPR-cas systems according to the core-gene tree is supported by Fig. S7, which shows a maximum-likelihood tree of the cas 1 gene for type-II CRISPR-cas, providing evidence of CRISPR-cas systems being acquired and maintained in the common ancestors of these sub-clades. The 6 strains with no CRISPRs show some clustering on the core-gene tree in Fig. 2, but JCM1045 and DSM18933 are singletons. The absence of CRISPR-cas systems does not have an obvious association with niche or the presence of prophages, suggesting that the interaction between CRISPR-cas systems, bacteriophages and the environment is far from straightforward. There are also 6 undefined CRISPRs from 4 strains that could not be described due to the absence of cas genes in close proximity. These CRISPRs are probably degraded systems that are no longer functional since all functioning CRISPR-cas systems have the cas 1 gene, which is involved in recognition and cleavage of invading DNA.

Fig. S8 shows a heatmap of gene counts for insertion sequences across the 43 strains, divided up into their respective replicons. The most striking thing about this figure is the inter-strain diversity of transposases, both within and between replicons. The gene counts for each transposase family in a specific strain on a particular replicon range from 0 to 52, highlighting the considerable variation in copy number displayed by these horizontally transferred sequences. The majority of transposases have copies on the chromosome and the plasmids, suggesting that they utilise the conjugative ability of plasmids to increase their abundance within and between species. The distributions of the IS families follow different patterns, from being widely spread over all three replicon groups (IS3) to being

limited to the chromosome and megaplasmid (IS21) to being confined to the smaller plasmid(s) (IS256). The only IS family confined to the chromosome is IS1 in a single strain - NIAS840.

The multi-replicon distribution of IS families implies that there is strong selection pressure on insertion sequences to transpose regularly from chromosomes to plasmids and vice versa, perhaps being partly responsible for the fact that transposases are currently considered to account for the most abundant gene families in both prokaryotes and eukaryotes [30]. The widespread distribution of IS3 in *L. salivarius* replicons is mirrored by its abundance (539 genes across the 43 strains); it is also the only family to consist of two sub-families - IS3 and IS150. Similarly, the other IS families with the widest distributions - ISL3, IS21 and IS200 - also have the greatest abundances after IS3, although IS21 is absent from the smaller plasmids even if it is ubiquitous on the *L. salivarius* chromosome.

Overall, IS families with a higher copy number in this dataset show a strong correlation with how many strains (and replicons per strain) harbour them (Spearman; rho = 0.95; $p < 0.001$), showing that insertion sequences do not have a tendency to just replicate within a single replicon without undergoing regular HGT. Out of the 19 IS families present in the TnpPred database (http://www.mobilomics.cl/), 14 are identified in *L. salivarius* in this study. This emphasises the ability of transposases to transfer themselves within and between species, leading to greater sequence diversity and, when they carry additional genes with them, greater functional diversity as well.

Protein secretion and membrane-anchoring gene richness are not associated with strain isolation source

Fig. S9 shows a barplot with the number of genes containing signal peptides and trans-membrane domains in the 43 strains. Proteins belonging to these two functional groups play an important role in the interaction of a bacterium with its environment since they are either secreted from the cell or function as membrane-bound structures. The number of predicted genes with signal peptides and with trans-membrane domains range from 56 to 84 and from 33 to 54, respectively. There is no association between niche and the number of either of these functional groups, which is not entirely surprising. These results highlight once more the point made earlier that certain isolation sources of *L. salivarius* strains shouldn't be interpreted as the niches that each strain has adapted to over time - some strains might be acting as opportunists that don't persist in a given environment for long such as the *Lactobacillus* species from a 2007 study (mainly *L. rhamnosus*) that were isolated from the blood, cerebrospinal fluid, peritoneal fluid and intestinal fistula of immuno-compromised children [31].

Most *L. salivarius* strains harbour genes for two bile salt hydrolases

Fig. S10 shows a heatmap of BLAST scores for all the predicted bile salt hydrolase (Bsh) genes in the 42 *L. salivarius* strains. The ability to hydrolyse bile salts is a necessary trait for any bacterium that is adapted to traversing the initial sections of the gastro-intestinal tract in order to colonise the intestine. It is also a required function for probiotics since a potential probiotic without the ability to reach its target area (usually the colon) will be ineffective. All 42 *L. salivarius* strains have at least one Bsh gene while *L. hayakitensis* DSM18933 has none, suggesting that the common ancestor of *L. salivarius* and *L. hayakitensis* did not possess a Bsh gene, although it is possible that another strain of *L. hayakitensis* does harbour one or more; if this is the case then gene decay of the Bsh gene in DSM18933 is a likely explanation. Two Bsh genes - one on the chromosome and one on the megaplasmid - seems to be the typical organisation in *L. salivarius* as described by Claesson *et al*

[32] since 36 out of 42 strains fit this description. Four strains - CECT5713, JCM1230, LMG14476 and LMG14477 - have a single Bsh gene located on the chromosome while 2 strains - cp400 and JCM1046 - have three BSH genes, both having two on the megaplasmid and one on the chromosome.

The presence of at least one Bsh in all 42 *L. salivarius* strains reinforces the point that this species is commonly isolated from the GIT of humans and animals. The variable number of Bsh genes and their presence on both the chromosome and the megaplasmid suggests that there is variability in bile resistance across the strains. This was shown in a study by Fang *et al*, but they cautioned that bile resistance is independent of the bsh1 allele type (the Bsh on the megaplasmid of most strains) and they go on to show that, upon exposure to bile and cholate, a transcriptome analysis reveals the up-regulation of numerous stress response and efflux proteins, which might mask the variable influence of Bsh allele types [33].

It should be noted that for the BLAST analysis of this category, a stricter cut-off value of 50% for coverage of both the query and reference genes was used. This was done because the number of BLAST hits to Bsh genes in the database contradicted previous literature so a closer agreement in protein length between query and reference sequences was enforced. It is possible that large discrepancies between the lengths of sequences in the database and sequences in the predicted *L. salivarius* gene repertoire led to false negative Bsh predictions. When the criteria are relaxed to include only 50% coverage of the query gene (and not the reference) an extra Bsh is predicted in some strains and these might actually be genuine Bsh genes that this study has excluded.

Summary survey of virulence factors and antibiotic resistance genes

Fig. S11 shows a barplot of the predicted number of putative antibiotic resistance genes (AR) and virulence factors (VF) across the 43 strains. VFs range from 2 to 3 genes and ARs range from 7 to 16. Virulence and antibiotic resistance are two traits that are screened for when assessing the suitability of a strain to act as a probiotic [34] and these traits are particularly dangerous in clinical settings. Table S8 and Table S9 give a more detailed summary of these results for ARs and VFs, respectively, while data file 7 and data file 8 give the corresponding amino acid sequences in FASTA format (figshare; Data Bibliography of main text).

The most commonly predicted function for AR genes in this dataset is transport, specifically a subset of efflux pumps for such antibiotics as tetracycline, elfamycin, bacitracin, clindamycin, fosfomycin, dalfopristin and others. Efflux pumps evolved long before the advent of antibiotic usage in modern medicine and probably originated as a defence against toxic substances entering the cell [35] – a strategy that has more recently been used to confer antibiotic resistance to microbes from multiple drugs, leading to a health crisis in the effective treatment of infection with antibiotics.

Virulence factor identification depends very much on context; a probiotic trait in one setting can be labelled as a virulence factor in another - for instance, when a pathogen acquires the ability to survive intestinal transit in order to colonise the human colon. The most commonly predicted functions for VF genes in our dataset are for an ATP-dependent protease and a UDP-glucose pyrophosphorylase. Overall, there is a wide variety of functions for these potential VFs, both in the VF database and in the predicted functions for *L. salivarius*. This highlights the ongoing evolutionary competition between hosts and microbes, the defensive and counter-defensive adaptive traits that arise from unrelated proteins with an overlapping strategy – to evade host mechanisms and successfully colonise the host environment.

REFERENCES

1. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. Basic local alignment search tool. *Journal of molecular biology*. 1990 Oct 5;215(3):403-10. PubMed PMID: 2231712. Epub 1990/10/05. eng.
2. **Olsen GJ, Overbeek R, Larsen N, Marsh TL, McCaughey MJ, Maciukenas MA, *et al***. The Ribosomal Database Project. *Nucleic acids research*. 1992 May 11;20 Suppl:2199-200. PubMed PMID: 1598241. Pubmed Central PMCID: PMC333993. Epub 1992/05/11. eng.
3. **Wu D, Jospin G, Eisen JA**. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PloS one*. 2013;8(10):e77033.
4. **Wood DE, Salzberg SL**. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 2014;15(3):R46.
5. **Li Y, Canchaya C, Fang F, Raftis E, Ryan KA, van Pijkeren JP, *et al***. Distribution of megaplasmids in *Lactobacillus salivarius* and other lactobacilli. *Journal of bacteriology*. 2007 Sep;189(17):6128-39. PubMed PMID: 17586640. Pubmed Central PMCID: PMC1951925. Epub 2007/06/26. eng.
6. **Finn RD, Clements J, Eddy SR**. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*. 2011 Jul 1;39(Web Server issue):W29-37. PubMed PMID: 21593126.
7. **Haft DH, Selengut JD, White O**. The TIGRFAMs database of protein families. *Nucleic acids research*. 2003 Jan 1;31(1):371-3. PubMed PMID: 12520025.
8. **Plyusnin I, Holm L, Kankainen M**. LOCP—locating pilus operons in Gram-positive bacteria. *Bioinformatics* (Oxford, England). 2009 May 1, 2009;25(9):1187-8.
9. **Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y**. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research*. 2012 Jul;40(Web Server issue):W445-51. PubMed PMID: 22645317.
10. **van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP**. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic acids research*. 2013 Jul;41(Web Server issue):W448-53. PubMed PMID: 23677608. Pubmed Central PMCID: PMC3692055. Epub 2013/05/17. eng.
11. **Nielsen H, Engelbrecht J, Brunak S, von Heijne G**. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*. 1997 January 1, 1997;10(1):1-6.
12. **Krogh A, Larsson B, von Heijne G, Sonnhammer EL**. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*. 2001 Jan 19;305(3):567-80. PubMed PMID: 11152613. Epub 2001/01/12. eng.
13. **McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, *et al***. The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*. 2013 July 1, 2013;57(7):3348-57.
14. **Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, *et al***. VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*. 2005 Jan 1;33(Database Issue):D325-8. PubMed PMID: 15608208.
15. **Roux S, Enault F, Hurwitz BL, Sullivan MB**. VirSorter: mining viral signal from microbial genomic data. *PeerJ*. 2015;3:e985. PubMed PMID: 26038737. Pubmed Central PMCID: PMC4451026. Epub 2015/06/04. eng.
16. **Riadi G, Medina-Moenne C, Holmes DS**. TnpPred: A Web Service for the Robust Prediction of Prokaryotic Transposases. *Comp Funct Genomics*. 2012;2012:5.
17. **R Core Team. R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2015.
18. **Culhane AC, Thioulouse J, Perriere G, Higgins DG**. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* (Oxford, England). 2005 Jun 1;21(11):2789-90. PubMed PMID: 15797915. Epub 2005/03/31. eng.

19. **Charif D, Lobry JR**. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 207-32.
20. **Qin QL, Xie BB, Zhang XY, Chen XL, Zhou BC, Zhou J, *et al***. A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of bacteriology*. 2014 Jun;196(12):2210-5. PubMed PMID: 24706738. Pubmed Central PMCID: PMC4054180. Epub 2014/04/08. eng.
21. **Winker K**. Reuniting Phenotype and Genotype in Biodiversity Research. *BioScience*. 2009 September 1, 2009;59(8):657-65.
22. **Sun Z, Harris HM, McCann A, Guo C, Argimon S, Zhang W, *et al***. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nature commun*. 2015;6:8322. PubMed PMID: 26415554. Pubmed Central PMCID: PMC4667430. Epub 2015/09/30. eng.
23. **Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, *et al***. Explaining microbial population genomics through phage predation. *Nat Rev Micro*. 2009 11//print;7(11):828-36.
24. **Horvath P, Barrangou R**. CRISPR/Cas, the immune system of bacteria and archaea. *Science* (New York, NY). 2010 Jan 8;327(5962):167-70. PubMed PMID: 20056882. Epub 2010/01/09. eng.
25. **Ochman H, Lawrence JG, Groisman EA**. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000 05/18/print;405(6784):299-304.
26. **Canchaya C, Fournous G, Brussow H**. The impact of prophages on bacterial chromosomes. *Molecular microbiology*. 2004 Jul;53(1):9-18. PubMed PMID: 15225299. Epub 2004/07/01. eng.
27. **Edlund A, Santiago-Rodriguez TM, Boehm TK, Pride DT**. Bacteriophage and their potential roles in the human oral cavity. *J Oral Microbiol*. 2015;7:27423. PubMed PMID: 25861745. Pubmed Central PMCID: PMC4393417. Epub 2015/04/12. eng.
28. **Hitch G, Pratten J, Taylor PW**. Isolation of bacteriophages from the oral cavity. *Lett Appl Microbiol*. 2004;39(2):215-9. PubMed PMID: 15242464. Epub 2004/07/10. eng.
29. **Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, White RA, III, *et al***. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J*. 2012 05//print;6(5):915-26.
30. **Aziz RK, Breitbart M, Edwards RA**. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic acids research*. 2010 Jul;38(13):4207-17. PubMed PMID: 20215432. Pubmed Central PMCID: PMC2910039. Epub 2010/03/11. eng.
31. **Muszynski Z, Mirska I, Matuska K**. [*Lactobacillus* species as opportunistic pathogens in children]. *Przeglad epidemiologiczny*. 2007;61(1):79-84. PubMed PMID: 17702443. Epub 2007/08/19. Paleczki z rodzaju *Lactobacillus*--czynnik zakazen oportunistycznych u dzieci. pol.
32. **Claesson MJ, Li Y, Leahy S, Canchaya C, van Pijkeren JP, Cerdeno-Tarraga AM, *et al***. Multireplicon genome architecture of *Lactobacillus salivarius*. *Proc Natl Acad Sci USA*. 2006 Apr 25;103(17):6718-23. PubMed PMID: 16617113. Pubmed Central PMCID: PMC1436024. Epub 2006/04/18. eng.
33. **Fang F, Li Y, Bumann M, Raftis EJ, Casey PG, Cooney JC, *et al***. Allelic variation of bile salt hydrolase genes in *Lactobacillus salivarius* does not determine bile resistance levels. *Journal of bacteriology*. 2009 Sep;191(18):5743-57. PubMed PMID: 19592587. Pubmed Central PMCID: PMC2737978. Epub 2009/07/14. Eng.
34. **Bennedsen M, Stuer-Lauridsen B, Danielsen M, Johansen E**. Screening for Antimicrobial Resistance Genes and Virulence Factors via Genome Sequencing. *Appl Environ Microbiol*. 2011 Apr;77(8):2785-7. PubMed PMID: 21335393.
35. **Webber MA, Piddock LJV**. The importance of efflux pumps in bacterial antibiotic resistance. *J Antimicrob Chemother*. 2003 January 1, 2003;51(1):9-11.

Fig. S1

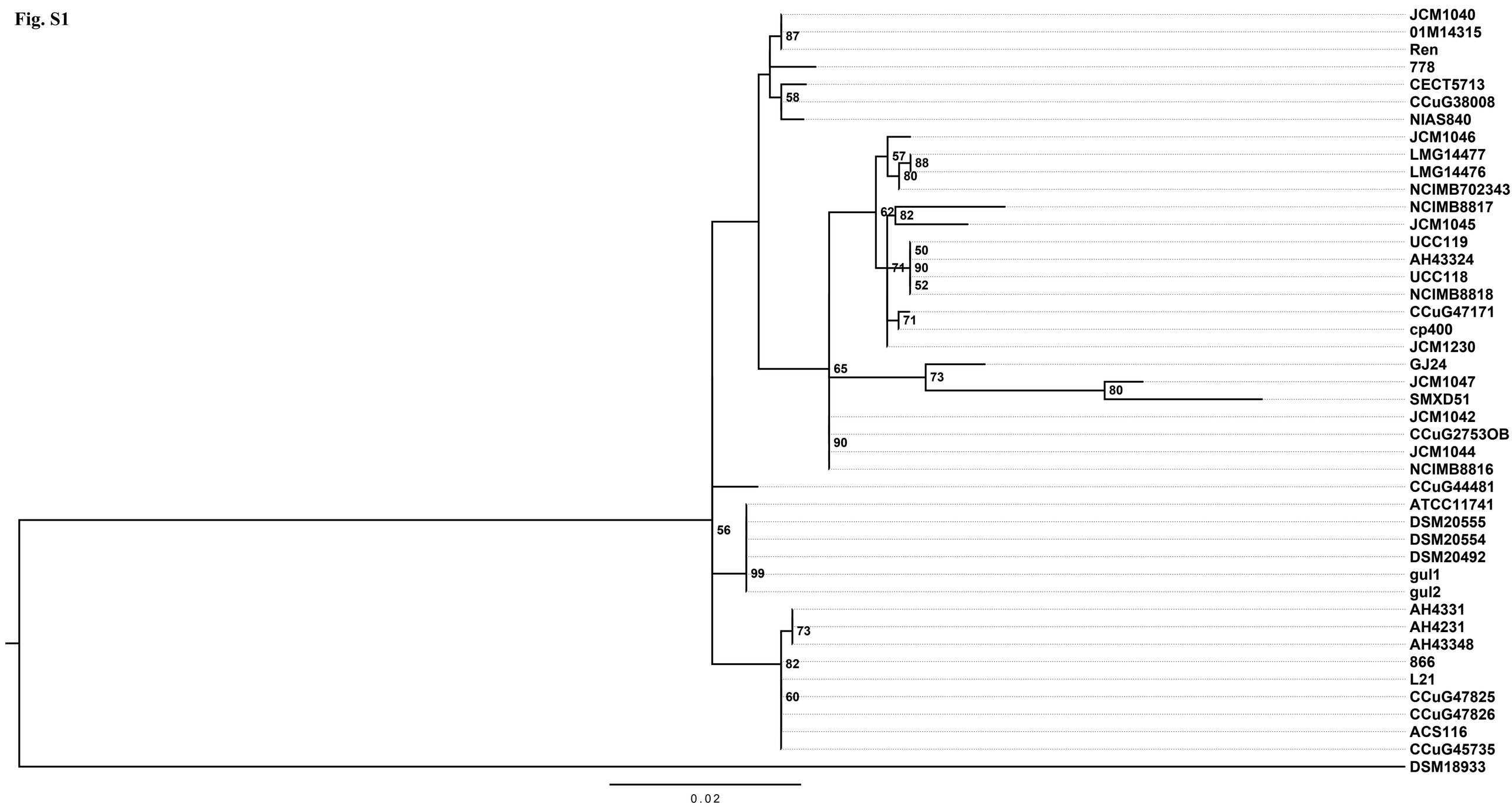

JCM1040
01M14315
Ren
778
CECT5713
CCuG38008
NIAS840
JCM1046
LMG14477
LMG14476
NCIMB702343
NCIMB8817
JCM1045
UCC119
AH43324
UCC118
NCIMB8818
CCuG47171
cp400
JCM1230
GJ24
JCM1047
SMXD51
JCM1042
CCuG2753OB
JCM1044
NCIMB8816
CCuG44481
ATCC11741
DSM20555
DSM20554
DSM20492
gul1
gul2
AH4331
AH4231
AH43348
866
L21
CCuG47825
CCuG47826
ACS116
CCuG45735
DSM18933
87
58
57
88
80
62
82
50
74
90
52
71
65
73
80
90
56
99
73
82
60
0.02

Fig. S2

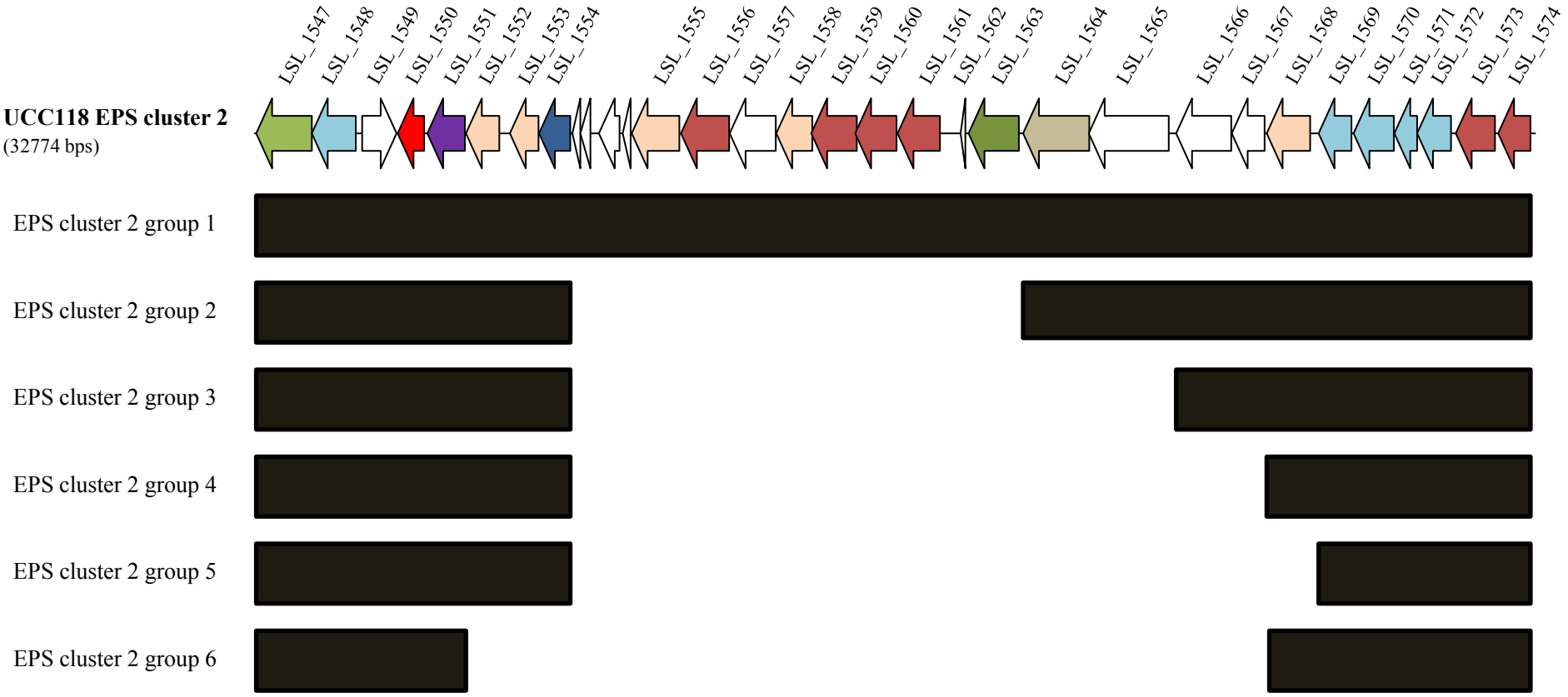

LSL_1547
LSL_1548
LSL_1549
LSL_1550
LSL_1551
LSL_1552
LSL_1553
LSL_1554
LSL_1555
LSL_1556
LSL_1557
LSL_1558
LSL_1559
LSL_1560
LSL_1561
LSL_1562
LSL_1563
LSL_1564
LSL_1565
LSL_1566
LSL_1567
LSL_1568
LSL_1569
LSL_1570
LSL_1571
LSL_1572
LSL_1573
LSL_1574
UCC118 EPS cluster 2
(32774 bps)
EPS cluster 2 group 1
EPS cluster 2 group 2
EPS cluster 2 group 3
EPS cluster 2 group 4
EPS cluster 2 group 5
EPS cluster 2 group 6

Fig. S3

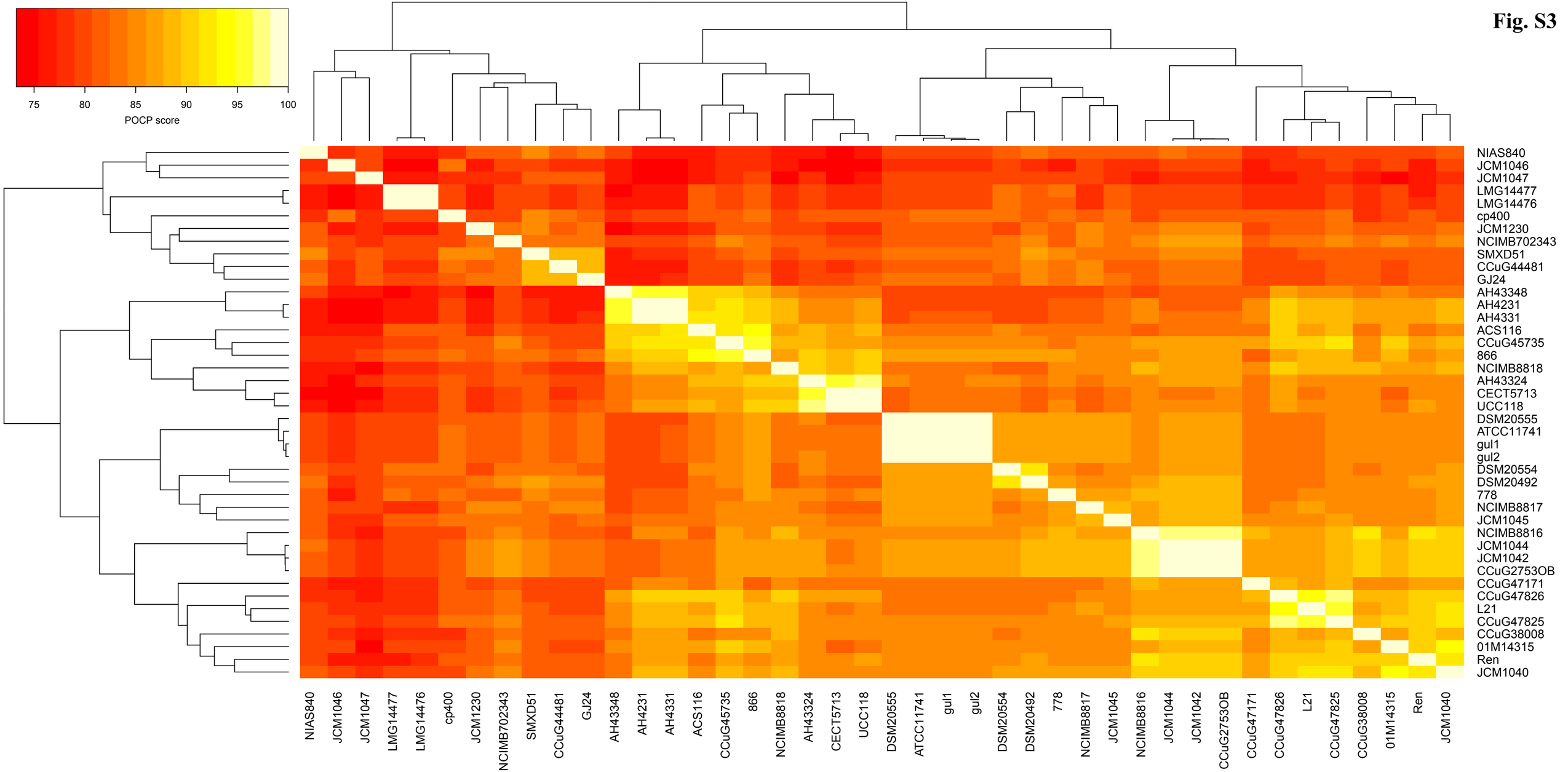

75
80
85
90
95
100
POCP score
NIAS840
JCM1046
JCM1047
LMG14477
LMG14476
cp400
JCM1230
NCIMB702343
SMXD51
CCuG44481
GJ24
AH43348
AH4231
AH4331
ACS116
CCuG45735
866
NCIMB8818
AH43324
CECT5713
UCC118
DSM20555
ATCC11741
gul1
gul2
DSM20554
DSM20492
778
NCIMB8817
JCM1045
NCIMB8816
JCM1044
JCM1042
CCuG2753OB
CCuG47171
CCuG47826
L21
CCuG47825
CCuG38008
01M14315
Ren
JCM1040

Fig. S4

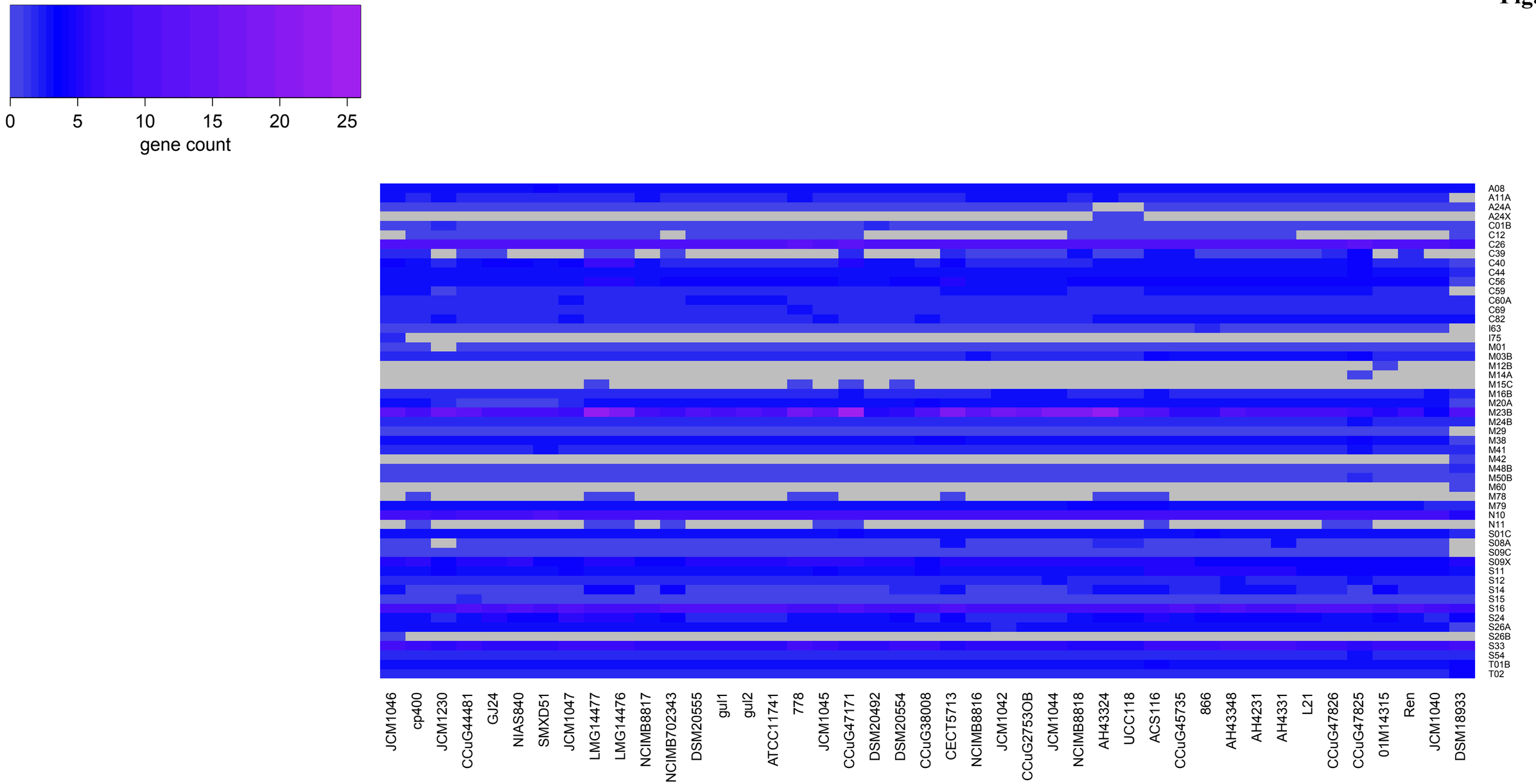
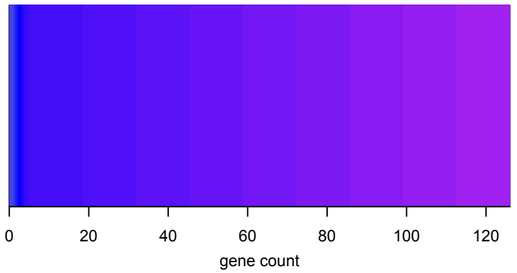

0
5
10
15
20
25
gene count
A08
A11A
A24A
A24X
C01B
C12
C26
C39
C40
C44
C56
C59
C60A
C69
C82
I63
I75
M01
M03B
M12B
M14A
M15C
M16B
M20A
M23B
M24B
M29
M38
M41
M42
M48B
M50B
M60
M78
M79
N10
N11
S01C
S08A
S09C
S09X
S11
S12
S14
S15
S16
S24
S26A
S26B
S33
S54
T01B
T02
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul1
gul2
ATCC11741
778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG2753OB
JCM1044
NCIMB8818
AH43324
UCC118
ACS116
CCuG45735
866
AH43348
AH4231
AH4331
L21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Fig. S5


0
20
40
60
80
100
120
gene count

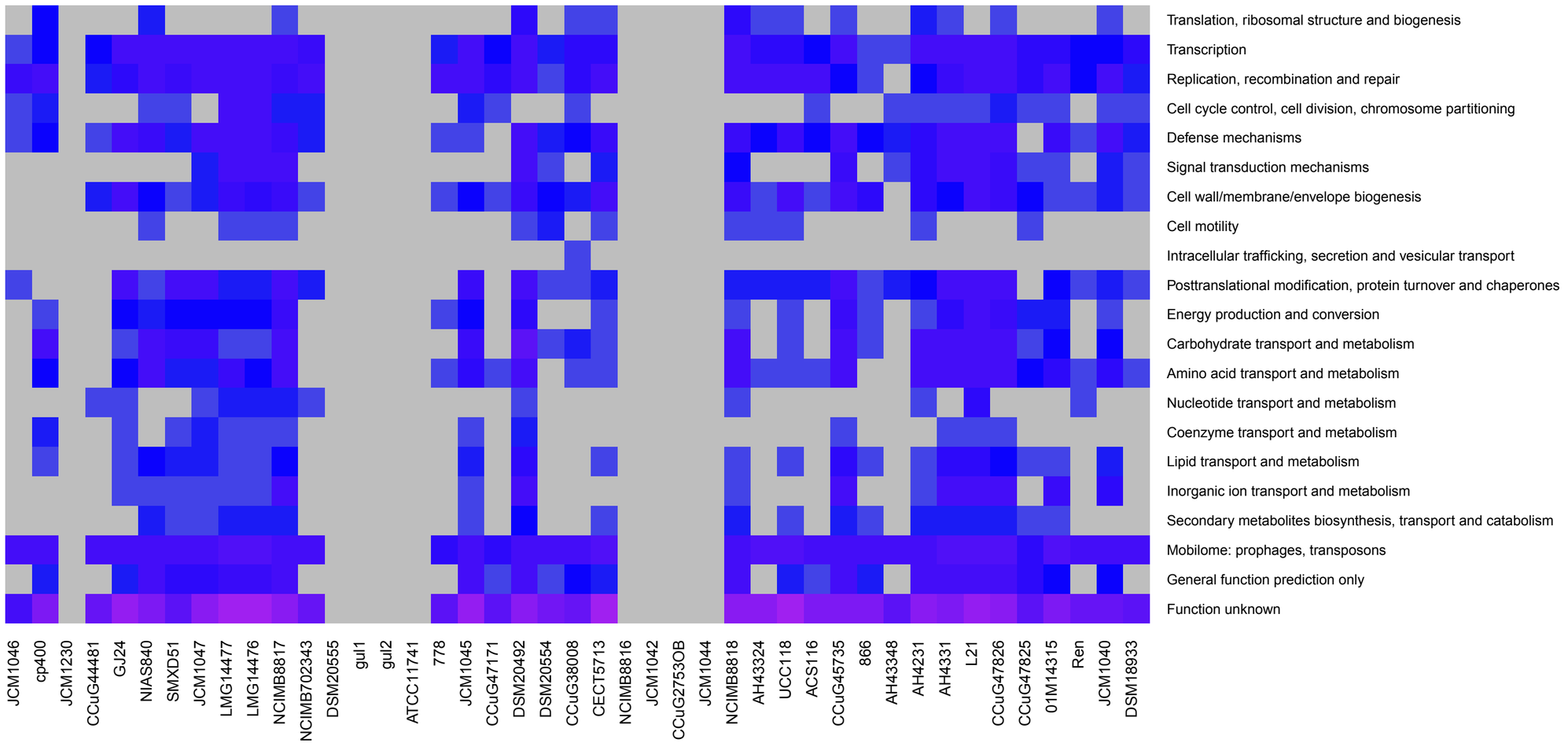

Translation, ribosomal structure and biogenesis
Transcription
Replication, recombination and repair
Cell cycle control, cell division, chromosome partitioning
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Intracellular trafficking, secretion and vesicular transport
Posttranslational modification, protein turnover and chaperones
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
Mobilome: prophages, transposons
General function prediction only
Function unknown
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul1
gul2
ATCC11741
778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG2753OB
JCM1044
NCIMB8818
AH43324
UCC118
ACS116
CCuG45735
866
AH43348
AH4231
AH4331
L21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Fig. S6

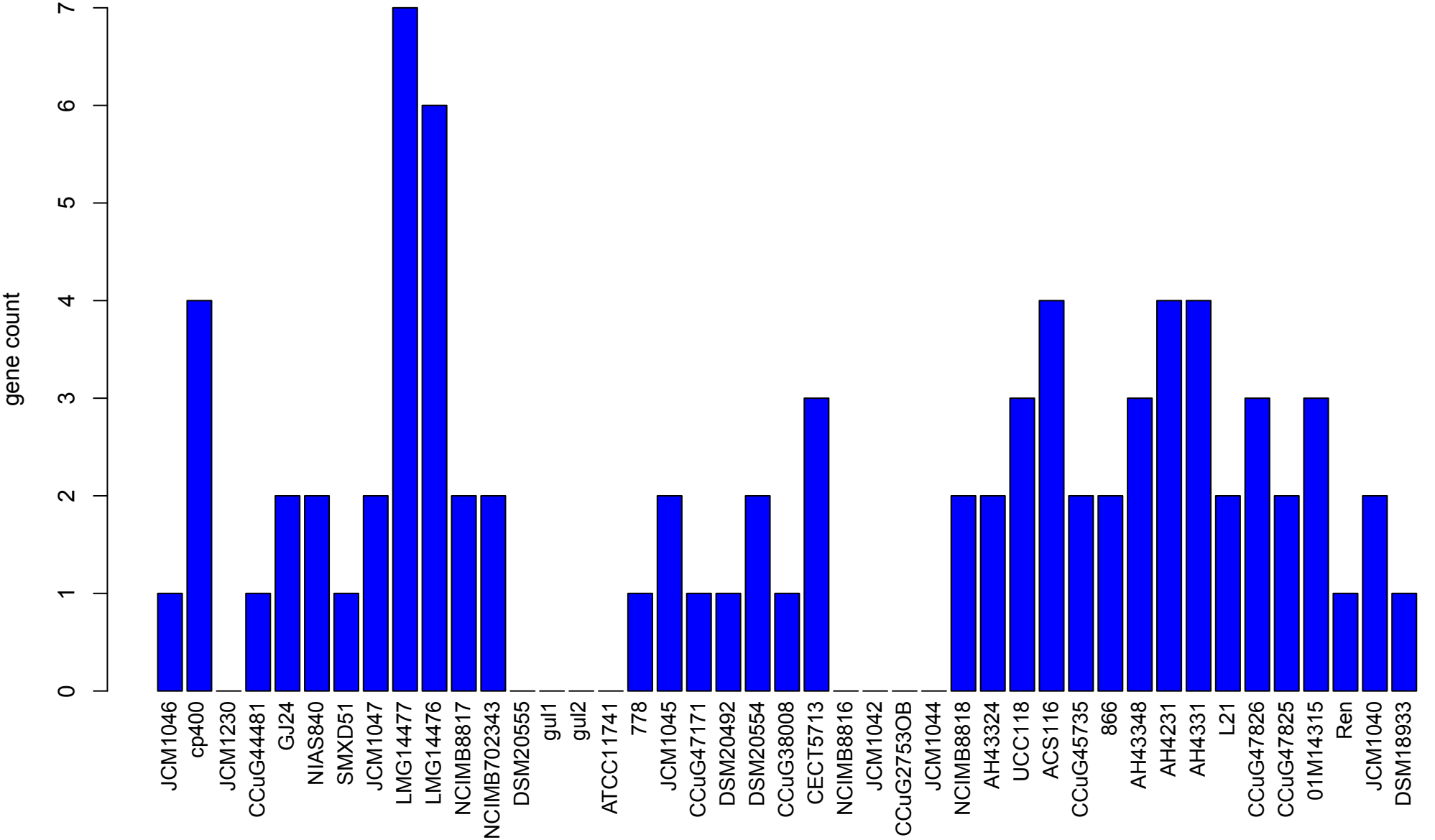

gene count
0
1
2
3
4
5
6
7
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul1
gul2
ATCC11741
778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG2753OB
JCM1044
NCIMB8818
AH43324
UCC118
ACS116
CCuG45735
866
AH43348
AH4231
AH4331
L21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Fig. S7

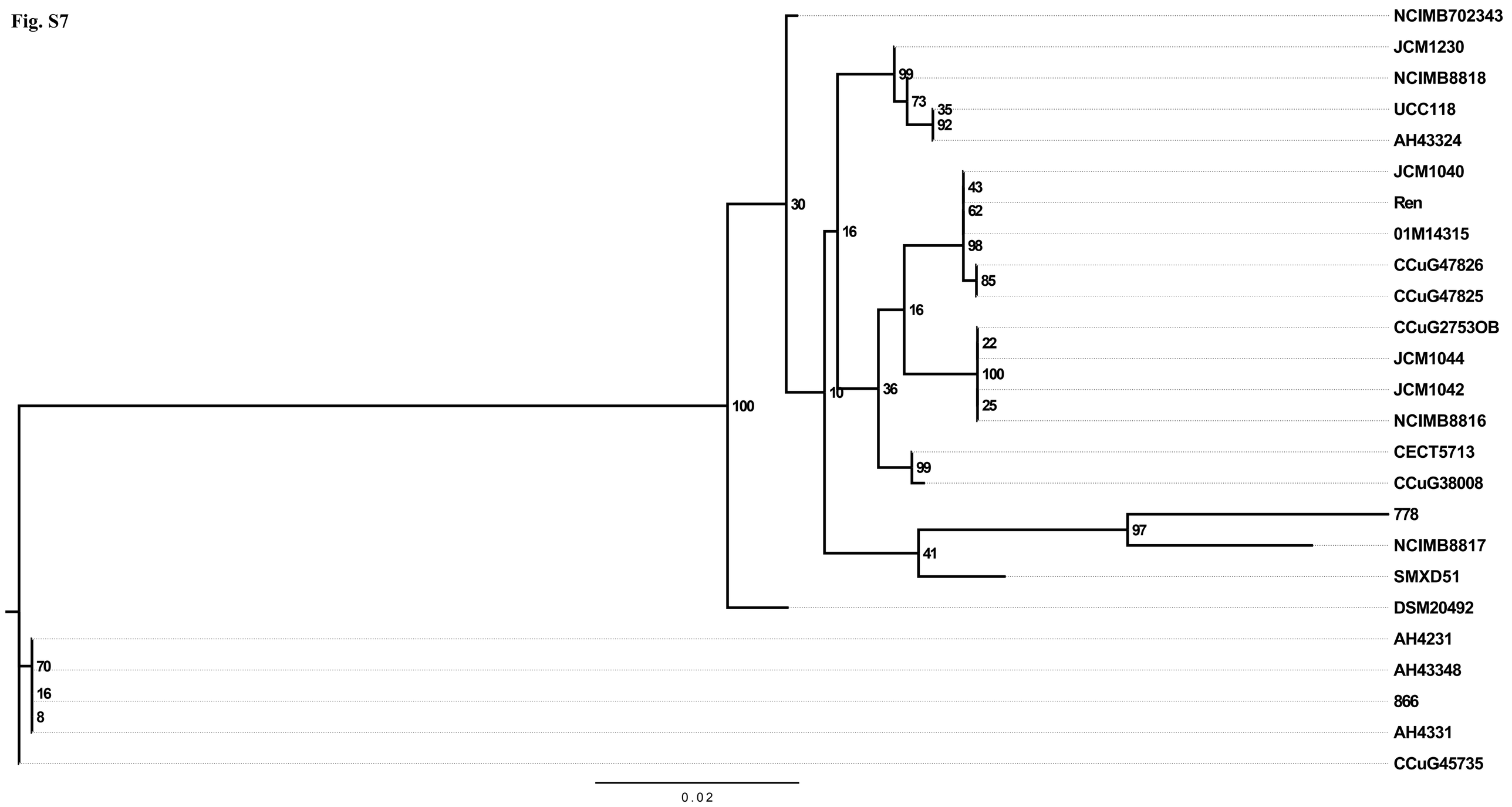

NCIMB702343
JCM1230
NCIMB8818
UCC118
AH43324
JCM1040
Ren
01M14315
CCuG47826
CCuG47825
CCuG2753OB
JCM1044
JCM1042
NCIMB8816
CECT5713
CCuG38008
778
NCIMB8817
SMXD51
DSM20492
AH4231
AH43348
866
AH4331
CCuG45735
99
73
35
92
43
62
98
85
16
22
100
25
30
16
10
36
99
97
41
100
70
16
8
0.02

Fig. S8

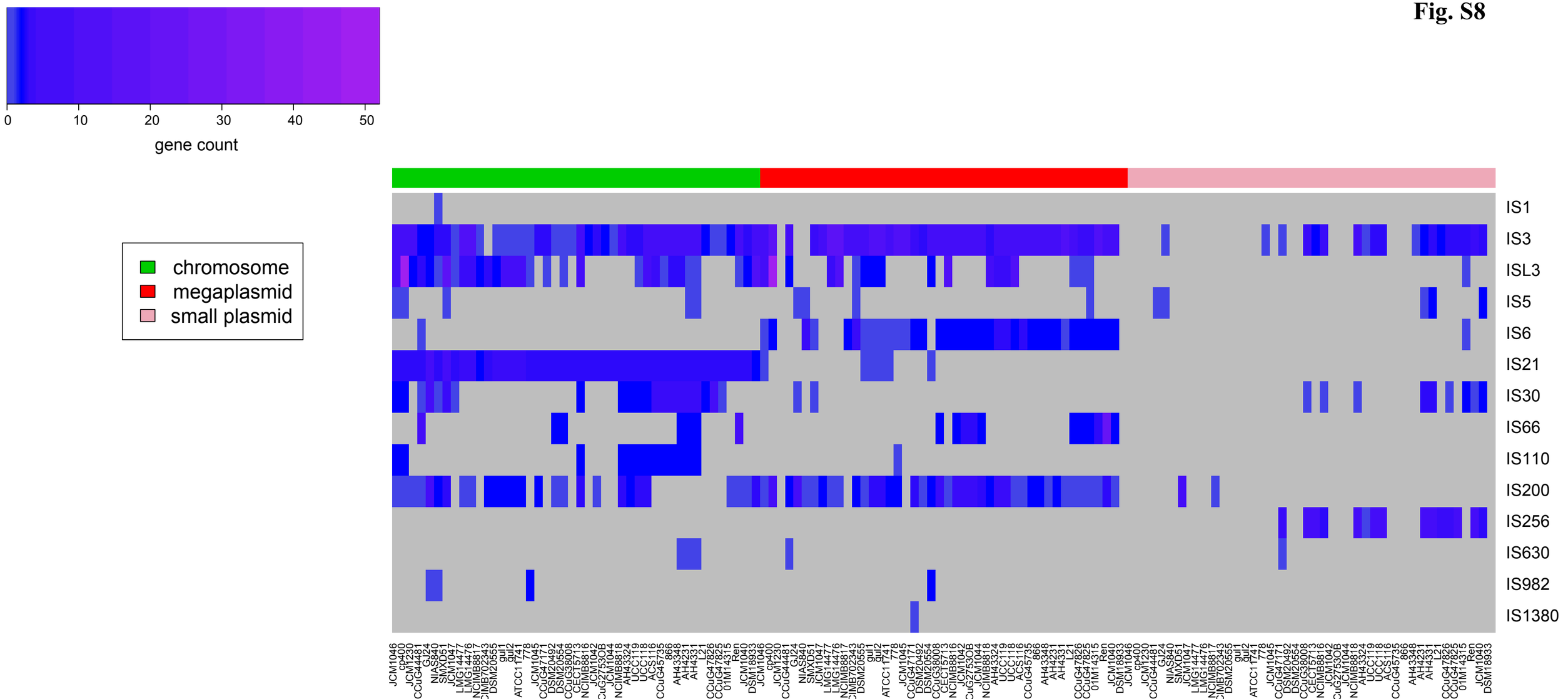

0
10
20
30
40
50
gene count
chromosome
megaplasmid
small plasmid
IS1
IS3
ISL3
IS5
IS6
IS21
IS30
IS66
IS110
IS200
IS256
IS630
IS982
IS1380
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul2
ATCC11778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG27530B
JCM1044
NCIMB8818
AH43324
UCC119
UCC118
ACS116
CCuG45735
866
AH43348
AH4331
21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul2
ATCC11778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG27530B
JCM1044
NCIMB8818
AH43324
UCC119
UCC118
ACS116
CCuG45735
866
AH43348
AH4331
21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul2
ATCC11778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG27530B
JCM1044
NCIMB8818
AH43324
UCC119
UCC118
ACS116
CCuG45735
866
AH43348
AH4331
21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Fig. S9

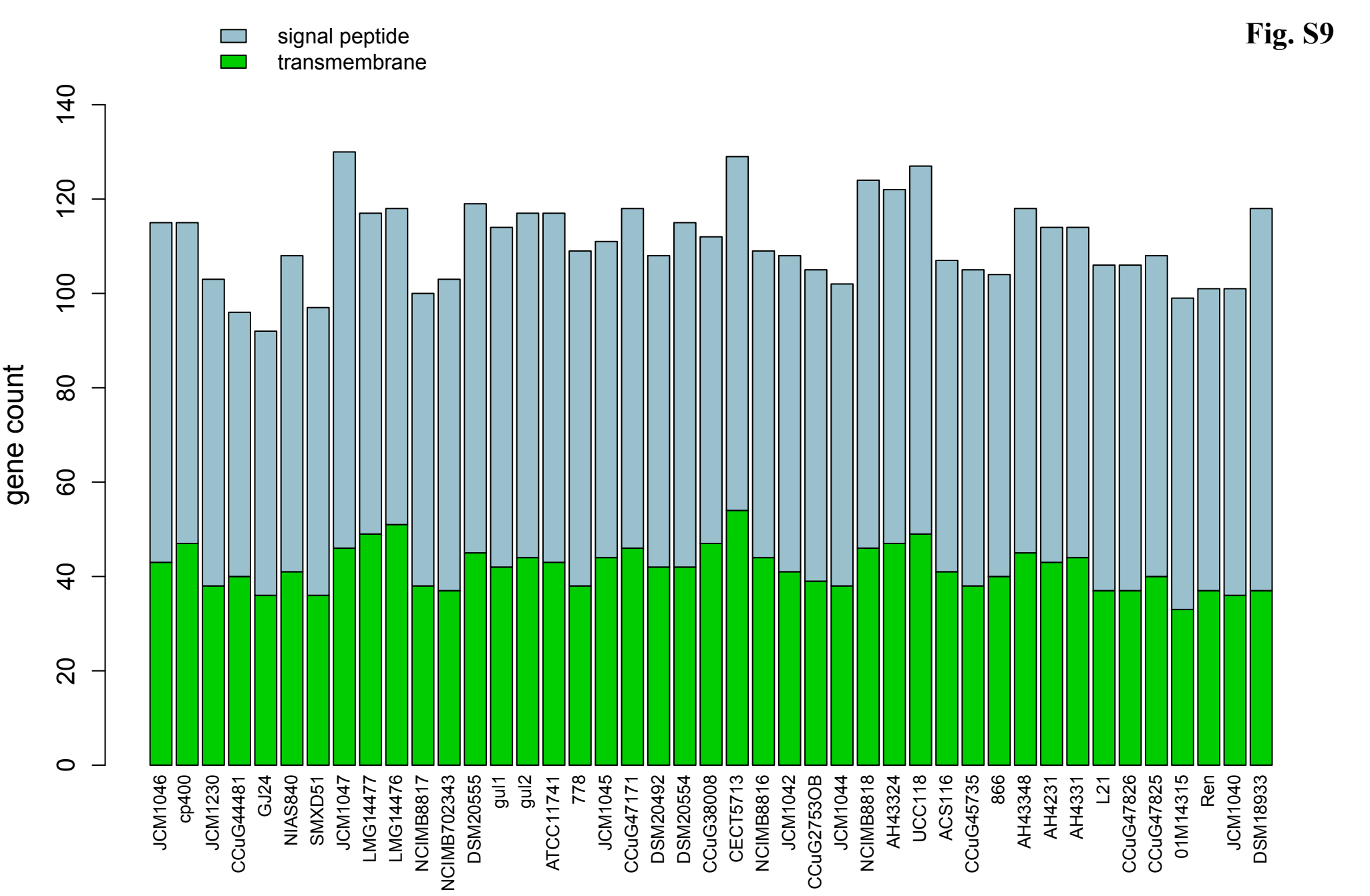

signal peptide
transmembrane
gene count
0
20
40
60
80
100
120
140
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul1
gul2
ATCC11741
778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG2753OB
JCM1044
NCIMB8818
AH43324
UCC118
ACS116
CCuG45735
866
AH43348
AH4231
AH4331
L21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Fig. S10

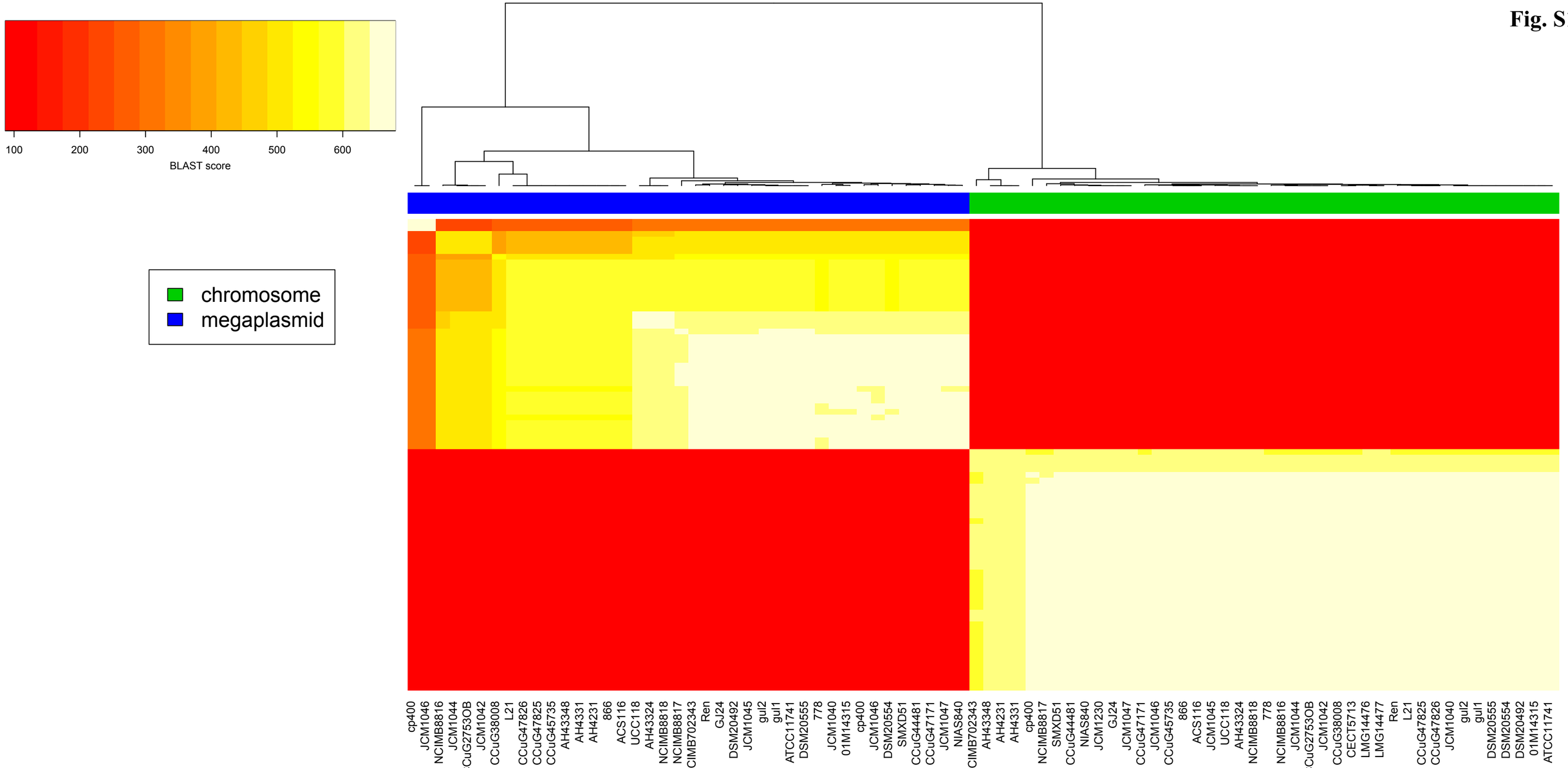

100
200
300
400
500
600
BLAST score
chromosome
megaplasmid
cp400
JCM1046
NCIMB8816
JCM1044
CCuG2753OB
JCM1042
CCuG38008
L21
CCuG47826
CCuG47825
CCuG45735
AH43348
AH4331
AH4231
866
ACS116
UCC118
AH43324
NCIMB8818
NCIMB8817
NCIMB702343
Ren
GJ24
DSM20492
JCM1045
gul2
gul1
ATCC11741
DSM20555
778
JCM1040
01M14315
cp400
JCM1046
DSM20554
SMXD51
CCuG44481
CCuG47171
JCM1047
NIAS840
NCIMB702343
AH43348
AH4231
AH4331
cp400
NCIMB8817
SMXD51
CCuG44481
NIAS840
JCM1230
GJ24
JCM1047
CCuG47171
JCM1046
CCuG45735
866
ACS116
JCM1045
UCC118
AH43324
NCIMB8818
778
NCIMB8816
JCM1044
CCuG2753OB
JCM1042
CCuG38008
CECT5713
LMG14476
LMG14477
Ren
L21
CCuG47825
CCuG47826
JCM1040
gul2
gul1
DSM20555
DSM20554
DSM20492
01M14315
ATCC11741

Fig. S11

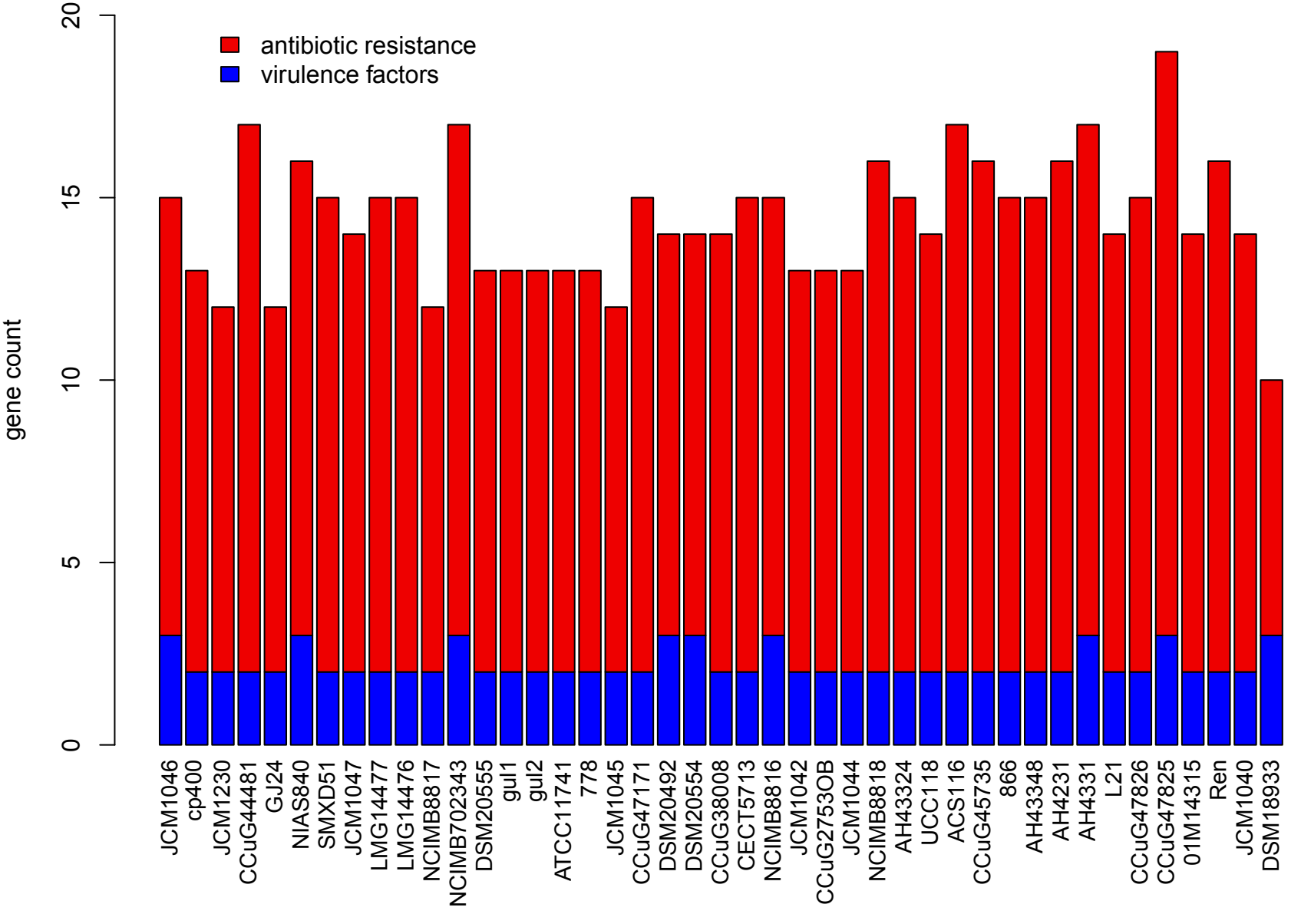

antibiotic resistance
virulence factors
gene count
0
5
10
15
20
JCM1046
cp400
JCM1230
CCuG44481
GJ24
NIAS840
SMXD51
JCM1047
LMG14477
LMG14476
NCIMB8817
NCIMB702343
DSM20555
gul1
gul2
ATCC11741
778
JCM1045
CCuG47171
DSM20492
DSM20554
CCuG38008
CECT5713
NCIMB8816
JCM1042
CCuG2753OB
JCM1044
NCIMB8818
AH43324
UCC118
ACS116
CCuG45735
866
AH43348
AH4231
AH4331
L21
CCuG47826
CCuG47825
01M14315
Ren
JCM1040
DSM18933

Table S1

	contigs	total size (bp)	max contig size (bp)	N50	GC %	gene count	circular megaplasmids	linear megaplasmids	small plasmids	isolation source	BioSample accession
ACS116	154	2044600	71067	28667	32.7	2210	1	0	0	human_vaginal_cavity	SAMN00017035
CCuG47825	120	1921003	93995	30969	32.8	2041	1	0	1	human_blood,55-year-old_female	SAMN06163258
LMG14477	243	2080357	139646	31335	32.8	2238	1	0	1	parakeet_with_sepsis	SAMN06163272
LMG14476	238	2087383	139606	38865	32.8	2243	1	0	1	cat_with_myocarditis	SAMN06163271
DSM18933	216	1699788	420945	47652	34	1844	0	0	0	faeces_of_thoroughbred_horse	SAMD00008721
cp400	89	2148103	152599	48732	32.8	2286	1	0	0	pre-weaned_piglet_faeces	SAMEA3138854
AH4331	177	2090404	196933	51266	32.9	2254	1	0	2	human_ileo-cecal_region	SAMN06163250
AH4231	186	2095546	196492	51450	32.9	2260	1	0	2	human_ileo-cecal_region	SAMN06163249
AH43324	114	2176341	196739	53856	32.8	2319	1	0	1	human_ileo-cecal_region	SAMN06163251
CCuG45735	77	1949679	199587	61326	32.7	1982	1	0	1	human_blood	SAMN06163256
CCuG44481	119	1946114	255195	63223	32.7	2013	1	0	1	bird	SAMN06163255
866	82	1973474	239143	67307	32.6	2019	1	0	0	clinical_isolate_ICU	SAMN03198074
AH43348	109	2026883	186083	73962	32.6	2177	1	1	1	human_ileo-cecal_region	SAMN06163252
01M14315	87	1933921	212516	78580	33	1965	1	0	1	human_gallbladder_pus	SAMN06163248
DSM20555	63	1982794	244105	80533	32.4	2000	1	0	0	human_saliva	SAMN02369414
L21	74	1956328	215606	95624	32.7	2015	1	0	1	human_faeces	SAMN06163270
NCIMB8816	77	1859216	465152	106164	32.8	1886	1	0	2	human_saliva	SAMN06163274
JCM1044	59	1800424	465193	115042	32.6	1799	1	0	0	human_intestine	SAMN06163266
JCM1042	60	1802147	465022	115144	32.6	1800	1	0	0	human_intestine	SAMN06163265
CCuG2753OB	48	1801346	465248	115144	32.6	1799	1	0	0	human_abdomen_abcess	SAMN06163253
NCIMB702343	83	1922903	646361	119088	32.8	1953	1	0	0	unknown	SAMN06163273
CCuG38008	91	1936298	463867	122107	32.7	1972	1	0	2	human_gall,73-year-old_male	SAMN06163254
gul-2	81	2002733	521962	126351	32.5	2027	1	0	0	root_canal	SAMN06163263
ATCC11741	54	1995868	274210	126392	32.5	2011	1	0	0	human_HMP_ref	SAMN00001483
JCM1040	66	1922028	421976	133271	32.8	1951	1	0	1	human_intestine	SAMN06163264
CCuG47171	140	2040146	353585	136925	32.9	2152	1	0	2	human_tooth_plaque	SAMN06163257
CCuG47826	71	1980366	279863	142710	32.8	2044	1	0	1	human_blood,55-year-old_female	SAMN06163259
JCM1230	82	1723361	400300	151438	32.6	1719	0	0	0	chicken_intestine	SAMN06163269
JCM1045	69	1928686	331541	153948	32.7	1964	1	0	0	human_intestine	SAMN06163267
gul-1	75	2001390	521857	162200	32.5	2026	1	0	0	root_canal	SAMN06163262
778	44	1942335	332399	164413	32.7	1947	1	0	1	clinical_isolate_ICU	SAMN03197988
DSM20554	74	1975060	334317	170258	32.6	2026	1	0	1	human_saliva	SAMN06163261
JCM1047	147	2222264	475807	178703	32.4	2345	1	1	1	swine_intestine	SAMN06163268
NCIMB8818	79	2013336	250622	180395	32.9	2103	1	0	1	St_Ivel_cheese	SAMN06163276
DSM20492	32	1889334	481851	240870	32.6	1892	1	0	0	human_saliva	SAMN06163260
NCIMB8817	56	1831814	777665	294813	32.6	1852	1	0	1	turkey_faeces	SAMN06163275
GJ24	11	1995968	754247	502388	33	2028	1	0	1	human_intestine	SAMN02470918
SMXD51	10	1967688	1019433	1019433	32.9	1992	1	0	2	chicken_cecum	SAMN02470767
NIAS840	4	2046557	1705688	1705688	33	2032	2	0	1	chicken_faeces	SAMN02470897
Ren	3	1978364	1751565	1751565	33	2019	1	0	1	human_centanarian_faeces	SAMN02584770
UCC118	4	2133977	1827111	1827111	33	2264	1	0	2	human_ileo-cecal_region	SAMN02604111
CECT5713	4	2136138	1828169	1828169	33.1	2345	1	0	2	human_breast_milk/infant_faeces	SAMN02604101
JCM1046	5	2320461	1836297	1836297	32.9	2296	2	1	1	swine_intestine	SAMN02711722

Table S2

strain	gene	contig	replicon	% identity	alignment length	e-value	BLAST score	gene length
778	repA_LSL1739	NODE_46	MP	99.21	254	8.00E-165	507	254
778	repE_LSL1740	NODE_46	MP	99.09	330	0	666	330
778	parA_LSL1741	NODE_46	MP	99.63	270	2.00E-179	550	270
866	repA_LSL1739	NODE_77	MP	99.21	254	8.00E-165	507	254
866	repE_LSL1740	NODE_77	MP	96.06	330	0	644	330
866	parA_LSL1741	NODE_77	MP	99.63	270	2.00E-179	550	270
01M14315	repA_LSL1739	contig_34	MP	99.21	254	7.00E-165	507	254
01M14315	repE_LSL1740	contig_34	MP	99.09	330	0	665	330
01M14315	parA_LSL1741	contig_34	MP	99.63	270	1.00E-179	550	270
ACS116	repA_LSL1739	contig_13	MP	98.82	254	1.00E-166	505	254
ACS116	repE_LSL1740	contig_13	MP	99.09	330	0	667	330
ACS116	parA_LSL1741	contig_139	MP	99.16	239	1.00E-169	484	270
AH4231	repA_LSL1739	contig_95	MP	99.21	254	8.00E-165	507	254
AH4231	repE_LSL1740	contig_95	MP	98.79	330	0	664	330
AH4231	parA_LSL1741	contig_95	MP	99.63	270	2.00E-179	550	270
AH4331	repA_LSL1739	contig_101	MP	99.21	254	8.00E-165	507	254
AH4331	repE_LSL1740	contig_101	MP	98.79	330	0	664	330
AH4331	parA_LSL1741	contig_101	MP	99.63	270	2.00E-179	550	270
AH43324	repA_LSL1739	contig_44	MP	99.61	254	5.00E-166	510	254
AH43324	repE_LSL1740	contig_44	MP	95.45	330	0	644	330
AH43324	parA_LSL1741	contig_44	MP	100	270	3.00E-180	551	270
AH43348	repA_LSL1739	contig_70	MP	99.21	254	9.00E-165	507	254
AH43348	repE_LSL1740	contig_70	MP	98.79	330	0	664	330
AH43348	parA_LSL1741	contig_70	MP	99.63	270	2.00E-179	550	270
AH43348	parA_LSL1741	contig_69	MPL	93.7	270	3.00E-169	520	270
ATCC11741	repA_LSL1739	contig_37	MP	99.21	254	9.00E-165	506	254
ATCC11741	repE_LSL1740	contig_37	MP	96.06	330	0	644	330
ATCC11741	parA_LSL1741	contig_37	MP	99.63	270	1.00E-179	550	270
CCuG2753OB	repA_LSL1739	contig_21	MP	99.21	254	7.00E-165	507	254
CCuG2753OB	repE_LSL1740	contig_21	MP	98.79	330	0	663	330
CCuG2753OB	parA_LSL1741	contig_21	MP	98.89	270	3.00E-178	546	270
CCuG38008	repA_LSL1739	contig_40	MP	99.21	254	5.00E-165	507	254
CCuG38008	repE_LSL1740	contig_40	MP	97.27	330	0	659	330
CCuG38008	parA_LSL1741	contig_40	MP	99.63	270	1.00E-179	550	270
CCuG44481	repA_LSL1739	contig_46	MP	99.21	254	7.00E-165	507	254
CCuG44481	repE_LSL1740	contig_46	MP	96.36	330	0	646	330
CCuG44481	parA_LSL1741	contig_39	MP	99.63	270	0	550	270
CCuG45735	repA_LSL1739	contig_51	MP	99.21	254	8.00E-165	507	254
CCuG45735	repE_LSL1740	contig_51	MP	99.39	330	0	668	330
CCuG45735	parA_LSL1741	contig_51	MP	99.63	270	2.00E-179	550	270
CCuG47171	repA_LSL1739	contig_58	MP	98.43	254	3.00E-163	503	254
CCuG47171	repE_LSL1740	contig_58	MP	96.97	330	0	650	330
CCuG47171	parA_LSL1741	contig_58	MP	100	270	4.00E-180	551	270
CCuG47825	repA_LSL1739	contig_92	MP	99.21	254	7.00E-165	507	254
CCuG47825	repE_LSL1740	contig_92	MP	99.39	330	0	668	330
CCuG47825	parA_LSL1741	contig_91	MP	99.63	270	0	550	270
CCuG47826	repA_LSL1739	contig_36	MP	99.21	254	8.00E-165	507	254
CCuG47826	repE_LSL1740	contig_36	MP	99.39	330	0	668	330
CCuG47826	parA_LSL1741	contig_36	MP	99.63	270	2.00E-179	550	270
CECT5713	repA_LSL1739	contig_4	MP	99.21	254	9.00E-165	507	254
CECT5713	repE_LSL1740	contig_4	MP	97.27	330	0	659	330
CECT5713	parA_LSL1741	contig_4	MP	99.63	270	2.00E-179	550	270
cp400	repA_LSL1739	contig_16	MP	99.21	254	7.00E-165	507	254
cp400	repE_LSL1740	contig_16	MP	96.36	330	0	646	330
cp400	parA_LSL1741	contig_16	MP	99.63	270	2.00E-179	550	270
DSM18933	repA_LSL1739	NODE_63	C	78.74	254	3.00E-132	413	254
DSM18933	parA_LSL1741	NODE_63	C	86.62	269	8.00E-157	484	270
DSM20492	repA_LSL1739	contig_14	MP	98.82	254	3.00E-164	505	254
DSM20492	repE_LSL1740	contig_14	MP	95.76	330	0	644	330
DSM20492	parA_LSL1741	contig_14	MP	99.63	270	2.00E-179	550	270

DSM20554	repA_LSL1739	contig_26	MP	99.21	254	8.00E-165	507	254
DSM20554	repE_LSL1740	contig_26	MP	92.73	330	0	626	330
DSM20554	parA_LSL1741	contig_26	MP	99.26	270	1.00E-178	547	270
DSM20555	repA_LSL1739	Scaffold29	MP	99.21	254	6.00E-165	506	254
DSM20555	repE_LSL1740	Scaffold29	MP	96.06	330	0	644	330
DSM20555	parA_LSL1741	Scaffold33	MP	99.63	270	3.00E-180	550	270
GJ24	repA_LSL1739	contig_9	MP	99.61	254	2.00E-165	509	254
GJ24	repE_LSL1740	contig_9	MP	95.15	330	0	641	330
GJ24	parA_LSL1741	contig_9	MP	99.63	270	2.00E-179	550	270
gul1	repA_LSL1739	contig_17	MP	99.21	254	9.00E-165	506	254
gul1	repE_LSL1740	contig_17	MP	96.06	330	0	644	330
gul1	parA_LSL1741	contig_17	MP	99.63	270	1.00E-179	550	270
gul2	repA_LSL1739	contig_18	MP	99.21	254	9.00E-165	506	254
gul2	repE_LSL1740	contig_18	MP	96.06	330	0	644	330
gul2	parA_LSL1741	contig_18	MP	99.63	270	1.00E-179	550	270
JCM1040	repA_LSL1739	contig_30	MP	99.21	254	7.00E-165	507	254
JCM1040	repE_LSL1740	contig_30	MP	99.7	330	0	671	330
JCM1040	parA_LSL1741	contig_30	MP	99.26	270	7.00E-179	548	270
JCM1042	repA_LSL1739	contig_27	MP	99.21	254	7.00E-165	507	254
JCM1042	repE_LSL1740	contig_27	MP	98.79	330	0	663	330
JCM1042	parA_LSL1741	contig_27	MP	98.89	270	3.00E-178	546	270
JCM1044	repA_LSL1739	contig_23	MP	99.21	254	7.00E-165	507	254
JCM1044	repE_LSL1740	contig_23	MP	98.79	330	0	663	330
JCM1044	parA_LSL1741	contig_23	MP	98.89	270	3.00E-178	546	270
JCM1045	repA_LSL1739	contig_30	MP	99.21	254	8.00E-165	507	254
JCM1045	repE_LSL1740	contig_30	MP	99.39	330	0	668	330
JCM1045	parA_LSL1741	contig_30	MP	99.63	270	2.00E-179	550	270
JCM1046	repA_LSL1739	contig_65	MP	99.21	254	8.00E-166	507	254
JCM1046	repE_LSL1740	contig_65	MP	96.67	330	0	648	330
JCM1046	parA_LSL1741	contig_65	MP	99.63	270	0	550	270
JCM1047	repA_LSL1739	contig_34	MP	98.43	254	3.00E-163	503	254
JCM1047	repE_LSL1740	contig_34	MP	95.45	330	0	642	330
JCM1047	parA_LSL1741	contig_34	MP	99.63	270	2.00E-179	550	270
L21	repA_LSL1739	contig_44	MP	99.21	254	8.00E-165	507	254
L21	repE_LSL1740	contig_44	MP	99.39	330	0	668	330
L21	parA_LSL1741	contig_44	MP	99.63	270	2.00E-179	550	270
LMG14476	repA_LSL1739	contig_42	MP	98.82	254	3.00E-164	505	254
LMG14476	repE_LSL1740	contig_42	MP	95.15	330	0	640	330
LMG14476	parA_LSL1741	contig_42	MP	99.63	270	2.00E-179	550	270
LMG14477	repA_LSL1739	contig_41	MP	98.82	254	3.00E-164	505	254
LMG14477	repE_LSL1740	contig_41	MP	95.15	330	0	640	330
LMG14477	parA_LSL1741	contig_41	MP	99.63	270	2.00E-179	550	270
NCIMB702343	repA_LSL1739	NODE_62	MP	99.61	254	2.00E-165	509	254
NCIMB702343	repE_LSL1740	NODE_62	MP	95.45	330	0	644	330
NCIMB702343	parA_LSL1741	NODE_62	MP	99.63	270	2.00E-179	550	270
NCIMB8816	repA_LSL1739	contig_35	MP	99.21	254	7.00E-165	507	254
NCIMB8816	repE_LSL1740	contig_35	MP	98.79	330	0	663	330
NCIMB8816	parA_LSL1741	contig_35	MP	98.89	270	3.00E-178	546	270
NCIMB8817	repA_LSL1739	contig_19	MP	99.21	254	7.00E-165	507	254
NCIMB8817	repE_LSL1740	contig_19	MP	96.36	330	0	645	330
NCIMB8817	parA_LSL1741	contig_19	MP	99.63	270	2.00E-179	550	270
NCIMB8818	repA_LSL1739	contig_33	MP	99.21	254	8.00E-165	507	254
NCIMB8818	repE_LSL1740	contig_33	MP	99.39	330	0	668	330
NCIMB8818	parA_LSL1741	contig_33	MP	99.63	270	2.00E-179	550	270
NIAS840	repA_LSL1739	contig_3	MP	98.43	254	3.00E-163	503	254
NIAS840	repE_LSL1740	contig_3	MP	95.45	330	0	642	330
NIAS840	parA_LSL1741	contig_3	MP	100	270	4.00E-180	551	270
Ren	repA_LSL1739	plasmid_1	MP	99.21	254	8.00E-165	507	254
Ren	repE_LSL1740	plasmid_1	MP	99.09	330	0	665	330
Ren	parA_LSL1741	plasmid_1	MP	99.63	270	2.00E-179	550	270
SMXD51	repA_LSL1739	contig_7	MP	99.44	179	8.00E-140	357	254

SMXD51	repE_LSL1740	contig_7	MP	96.36	330	0	646	330
SMXD51	parA_LSL1741	contig_7	MP	99.63	270	2.00E-179	550	270
UCC118	repA_LSL1739	contig_4	MP	100	254	2.00E-166	511	254
UCC118	repE_LSL1740	contig_4	MP	95.45	330	0	644	330
UCC118	parA_LSL1741	contig_4	MP	100	270	5.00E-180	551	270

Table S3

strain	AOI	replicon	bacteriocin	class
778	AOI 1	chromosome	enterolysin A	III
778	AOI 1	chromosome	enterolysin A	III
866	AOI 1	chromosome	enterolysin A	III
866	AOI 1	chromosome	enterolysin A	III
866	AOI 1	megaplasmid	salivaricin P	II
01M14315	AOI 1	chromosome	enterolysin A	III
01M14315	AOI 1	chromosome	enterolysin A	III
ACS116	AOI 1	chromosome	enterolysin A	III
ACS116	AOI 1	chromosome	enterolysin A	III
ACS116	AOI 1	chromosome	enterolysin A	III
ACS116	AOI 1	chromosome	enterolysin A	III
ACS116	AOI 1	megaplasmid	salivaricin P	II
AH4231	AOI 1	chromosome	enterolysin A	III
AH4231	AOI 1	chromosome	enterolysin A	III
AH4231	AOI 1	chromosome	enterolysin A	III
AH4231	AOI 1	megaplasmid	salivaricin P	II
AH4331	AOI 1	chromosome	enterolysin A	III
AH4331	AOI 1	chromosome	enterolysin A	III
AH4331	AOI 1	chromosome	enterolysin A	III
AH4331	AOI 1	megaplasmid	salivaricin P	II
AH43324	AOI 1	chromosome	enterolysin A	III
AH43324	AOI 1	chromosome	enterolysin A	III
AH43324	AOI 1	chromosome	enterolysin A	III
AH43324	AOI 1	megaplasmid	salivaricin P	II
AH43348	AOI 1	chromosome	enterolysin A	III
AH43348	AOI 1	chromosome	enterolysin A	III
AH43348	AOI 1	chromosome	enterolysin A	III
AH43348	AOI 1	megaplasmid	salivaricin P	II
ATCC11741	AOI 1	chromosome	enterolysin A	III
CCUG2753OB	AOI 1	chromosome	enterolysin A	III
CCUG2753OB	AOI 1	megaplasmid	LS2	II
CCUG38008	AOI 1	chromosome	enterolysin A	III
CCUG38008	AOI 1	chromosome	enterolysin A	III
CCUG38008	AOI 1	megaplasmid	salivaricin P	II
CCUG44481	AOI 1	chromosome	enterolysin A	III
CCUG44481	AOI 1	megaplasmid	plantaricin S	II
CCUG44481	AOI 2	megaplasmid	plantaricin NC8	II
CCUG44481	AOI 2	megaplasmid	lactacin F	II
CCUG44481	AOI 2	megaplasmid	acidocin LF221B	II
CCUG44481	AOI 2	megaplasmid	salivaricin P	II
CCUG45735	AOI 1	chromosome	enterolysin A	III
CCUG45735	AOI 1	chromosome	enterolysin A	III
CCUG45735	AOI 1	megaplasmid	salivaricin P	II
CCUG47171	AOI 1	chromosome	enterolysin A	III
CCUG47171	AOI 1	chromosome	enterolysin A	III
CCUG47171	AOI 1	megaplasmid	plantaricin NC8	II
CCUG47171	AOI 1	megaplasmid	lactacin F	II
CCUG47171	AOI 1	megaplasmid	acidocin LF221B	II
CCUG47171	AOI 1	megaplasmid	salivaricin P	II

CCUG47825	AOI 1	chromosome	enterolysin A	III
CCUG47825	AOI 1	chromosome	enterolysin A	III
CCUG47826	AOI 1	chromosome	enterolysin A	III
CCUG47826	AOI 1	chromosome	enterolysin A	III
CCUG47826	AOI 1	chromosome	enterolysin A	III
CCUG47826	AOI 1	megaplasmid	salivaricin P	II
CECT5713	AOI 2	chromosome	enterolysin A	III
CECT5713	AOI 3	chromosome	enterolysin A	III
CECT5713	AOI 4	chromosome	enterolysin A	III
CECT5713	AOI 1	megaplasmid	salivaricin P	II
cp400	AOI 1	chromosome	enterolysin A	III
cp400	AOI 1	chromosome	enterolysin A	III
cp400	AOI 1	megaplasmid	salivaricin P	II
DSM18933	AOI 1	chromosome	enterolysin A	III
DSM18933	AOI 1	chromosome	enterolysin A	III
DSM20492	AOI 1	chromosome	enterolysin A	III
DSM20554	AOI 1	chromosome	enterolysin A	III
DSM20554	AOI 1	chromosome	enterolysin A	III
DSM20554	AOI 1	small plasmid	MR10B	II
DSM20555	AOI 1	chromosome	enterolysin A	III
GJ24	AOI 1	chromosome	enterolysin A	III
GJ24	AOI 1	megaplasmid	plantaricin S	II
GJ24	AOI 2	megaplasmid	plantaricin NC8	II
GJ24	AOI 2	megaplasmid	salivaricin P	II
gul1	AOI 1	chromosome	enterolysin A	III
gul2	AOI 1	chromosome	enterolysin A	III
JCM1040	AOI 1	chromosome	enterolysin A	III
JCM1040	AOI 1	chromosome	enterolysin A	III
JCM1042	AOI 1	chromosome	enterolysin A	III
JCM1042	AOI 1	megaplasmid	LS2	II
JCM1044	AOI 1	chromosome	enterolysin A	III
JCM1044	AOI 1	megaplasmid	LS2	II
JCM1045	AOI 1	chromosome	enterolysin A	III
JCM1045	AOI 1	megaplasmid	enterolysin A	III
JCM1046	AOI 1	chromosome	enterolysin A	III
JCM1046	AOI 1	megaplasmid	salivaricin P	II
JCM1046	AOI 1	linear megaplasmid	enterolysin A	III
JCM1046	AOI 1	small plasmid	MR10B	II
JCM1047	AOI 1	chromosome	enterolysin A	III
JCM1047	AOI 1	small plasmid	MR10B	II
JCM1047	AOI 1	megaplasmid	salivaricin P	II
JCM1047	AOI 1	linear megaplasmid	enterolysin A	III
JCM1230	AOI 1	chromosome	enterolysin A	III
L21	AOI 1	chromosome	enterolysin A	III
L21	AOI 1	chromosome	enterolysin A	III
L21	AOI 1	megaplasmid	salivaricin P	II
LMG14476	AOI 1	chromosome	enterolysin A	III
LMG14476	AOI 1	chromosome	enterolysin A	III
LMG14476	AOI 1	megaplasmid	salivaricin P	II
LMG14477	AOI 1	chromosome	enterolysin A	III

LMG14477	AOI 1	chromosome	enterolysin A	III
LMG14477	AOI 1	megaplasmid	salivaricin P	II
NCIMB702343	AOI 1	chromosome	enterolysin A	III
NCIMB702343	AOI 1	chromosome	enterolysin A	III
NCIMB702343	AOI 1	megaplasmid	salivaricin P	II
NCIMB8816	AOI 1	megaplasmid	LS2	II
NCIMB8816	AOI 1	chromosome	enterolysin A	III
NCIMB8817	AOI 1	chromosome	enterolysin A	III
NCIMB8818	AOI 1	chromosome	enterolysin A	III
NCIMB8818	AOI 1	chromosome	enterolysin A	III
NCIMB8818	AOI 1	megaplasmid	salivaricin P	II
NIAS840	AOI 1	chromosome	enterolysin A	III
Ren	AOI 1	chromosome	enterolysin A	III
Ren	AOI 2	chromosome	enterolysin A	III
SMXD51	AOI 1	chromosome	enterolysin A	III
SMXD51	AOI 1	megaplasmid	LS2	II
UCC118	AOI 1	chromosome	enterolysin A	III
UCC118	AOI 2	chromosome	enterolysin A	III
UCC118	AOI 3	chromosome	enterolysin A	III
UCC118	AOI 1	megaplasmid	salivaricin P	II

Table S4

strain	classification	taxon nucleotide %	taxon nucleotides	contig count
01M14315	Lactobacillus_salivarius	97.9012	1755411	67
01M14315	unclassified	1.96655	35261	6
01M14315	Lactobacillus_fermentum	0.124593	2234	1
01M14315	Lactobacillus_casei_group	0.00764064	137	1
778	Lactobacillus_salivarius	99.3133	1921066	36
778	unclassified	0.686742	13284	1
866	Lactobacillus_salivarius	100	1963667	75
ACS116	Lactobacillus_salivarius	100	1956854	132
AH4231	Lactobacillus_salivarius	97.1495	1950134	121
AH4231	unclassified	2.45587	49298	32
AH4231	Lactobacillus_casei_group	0.212867	4273	9
AH4231	Lactobacillus_fermentum	0.181732	3648	2
AH4331	Lactobacillus_salivarius	97.1714	1947771	115
AH4331	unclassified	2.47302	49571	33
AH4331	Lactobacillus_casei_group	0.181594	3640	8
AH4331	Lactobacillus_fermentum	0.174011	3488	1
AH43324	Lactobacillus_salivarius	98.4574	1942606	87
AH43324	unclassified	1.54259	30436	1
AH43348	Lactobacillus_salivarius	100	2028360	109
ATCC11741	Lactobacillus_salivarius	99.9018	1897727	40
ATCC11741	unclassified	0.0484841	921	4
ATCC11741	Corynebacterium_aurimucosum	0.0243736	463	2
ATCC11741	Erysipelothrix_rhusiopathiae	0.0126869	241	1
ATCC11741	Rhodococcus_equi	0.0126869	241	1
CCuG2753OB	Lactobacillus_salivarius	100	1790291	42
CCuG38008	Lactobacillus_salivarius	98.1458	1895829	67
CCuG38008	unclassified	1.82684	35288	10
CCuG38008	Lactobacillus_casei_group	0.0273342	528	2
CCuG44481	Lactobacillus_salivarius	96.0994	1812482	54
CCuG44481	unclassified	2.94823	55605	5
CCuG44481	Lactobacillus_reuteri	0.738793	13934	1
CCuG44481	Clostridium_sp._SY8519	0.213621	4029	1
CCuG45735	Lactobacillus_salivarius	99.9105	1947895	69
CCuG45735	Lactobacillus_helveticus	0.0802713	1565	1
CCuG45735	unclassified	0.00918119	179	1
CCuG47171	Lactobacillus_salivarius	97.6392	1974695	82
CCuG47171	unclassified	1.87664	37954	22
CCuG47171	Lactobacillus_casei_group	0.32352	6543	7
CCuG47171	Lactobacillus_plantarum	0.160598	3248	1
CCuG47825	Lactobacillus_salivarius	99.8319	1786862	97
CCuG47825	unclassified	0.114757	2054	3
CCuG47825	Streptococcus_dysgalactiae_group	0.0323487	579	2
CCuG47825	Citrobacter_rodentium_ICC168	0.0209512	375	1
CCuG47826	Lactobacillus_salivarius	99.1878	1940444	50
CCuG47826	unclassified	0.783915	15336	5
CCuG47826	Lactobacillus_sanfranciscensis	0.0221332	433	3
CCuG47826	Lactobacillus_casei_group	0.00618504	121	1
CECT5713	Lactobacillus_salivarius	100	2136138	4
cp400	Lactobacillus_salivarius	98.7922	2016437	72
cp400	Lactobacillus_reuteri	1.20779	24652	2
DSM18933	unclassified	93.4389	6964	18
DSM18933	Lactobacillus_salivarius	6.56112	489	3
DSM20492	Lactobacillus_salivarius	100	1890069	30
DSM20554	Lactobacillus_salivarius	100	1949872	51
DSM20555	Lactobacillus_salivarius	100	1874089	44

GJ24	Lactobacillus_salivarius	99.7466	1986988	8
GJ24	unclassified	0.253359	5047	1
gul1	Lactobacillus_salivarius	100	1907028	62
gul2	Lactobacillus_salivarius	99.982	1908848	68
gul2	unclassified	0.011628	222	1
gul2	Lactobacillus_johnsonii	0.00633776	121	1
JCM1040	Lactobacillus_salivarius	98.5093	1822228	46
JCM1040	unclassified	1.48416	27454	7
JCM1040	Lactobacillus_casei_group	0.00654124	121	1
JCM1042	Lactobacillus_salivarius	100	1789991	54
JCM1044	Lactobacillus_salivarius	100	1788967	53
JCM1045	Lactobacillus_salivarius	100	1928808	64
JCM1046	Lactobacillus_salivarius	94.2248	1802918	91
JCM1046	unclassified	4.7194	90302	6
JCM1046	Lactobacillus_reuteri	1.0558	20202	3
JCM1047	Lactobacillus_salivarius	98.613	2040949	81
JCM1047	unclassified	1.38704	28707	1
JCM1230	Lactobacillus_salivarius	100	1656507	38
L21	Lactobacillus_salivarius	99.7884	1927838	60
L21	unclassified	0.195349	3774	5
L21	Lactobacillus_sanfranciscensis	0.0162532	314	2
LMG14476	Lactobacillus_salivarius	100	1887993	157
LMG14477	Lactobacillus_salivarius	99.99	1822244	144
LMG14477	unclassified	0.00998669	182	1
NCIMB702343	Lactobacillus_salivarius	99.9607	1884140	42
NCIMB702343	unclassified	0.0393128	741	2
NCIMB8816	Lactobacillus_salivarius	99.0933	1820558	61
NCIMB8816	unclassified	0.785101	14424	7
NCIMB8816	Lactobacillus_casei_group	0.121597	2234	1
NCIMB8817	Lactobacillus_salivarius	100	1814802	45
NCIMB8818	Lactobacillus_salivarius	97.8194	1970529	67
NCIMB8818	unclassified	1.99051	40098	9
NCIMB8818	Lactobacillus_casei_group	0.120231	2422	2
NCIMB8818	Streptococcus_anginosus_group	0.0698948	1408	1
NIAS840	Lactobacillus_salivarius	100	1866462	3
Ren	Lactobacillus_salivarius	100	1978364	3
SMXD51	Lactobacillus_salivarius	98.4374	1928066	8
SMXD51	unclassified	1.56264	30607	1
UCC118	Lactobacillus_salivarius	100	2133977	4

Table S5

	UCC118 +	UCC118 -	CECT5713 +	CECT5713 -	NIAS840 +	NIAS840 -	Ren +	Ren -
chromosome	100	0	100	0	100	0	100	0
megaplasmid 1	100	0	100	0	97.5	2.5	85.3	14.7
megaplasmid 2	NA	NA	NA	NA	8.3	91.7	NA	NA
small plasmid 1	100	0	100	0	0	100	90.7	9.3
small plasmid 2	100	0	100	0	NA	NA	NA	NA

Table S6

Aspartic (A) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
A1	A1A	pepsin A (Homo sapiens)
	A1B	nepenthesin (Nepenthes gracilis)
A2	A2A	HIV-1 retropepsin (human immunodeficiency virus 1)
	A2B	Ty3 transposon peptidase (Saccharomyces cerevisiae)
	A2C	Gypsy transposon peptidase (Drosophila melanogaster)
	A2D	Osvaldo retrotransposon peptidase (Drosophila buzzatii)
A3	A3A	cauliflower mosaic virus-type peptidase (cauliflower mosaic virus)
	A3B	bacilliform virus peptidase (rice tungro bacilliform virus)
A5		thermopsin (Sulfolobus acidocaldarius)
A8		signal peptidase II (Escherichia coli)
A9		spumapepsin (human spumaretrovirus)
A11	A11A	Copia transposon peptidase (Drosophila melanogaster)
	A11B	Ty1 transposon peptidase (Saccharomyces cerevisiae)
A22	A22A	presenilin 1 (Homo sapiens)
	A22B	impas 1 peptidase (Homo sapiens)
A24	A24A	type 4 prepilin peptidase 1 (Pseudomonas aeruginosa)
	A24B	FlaK peptidase (Methanococcus maripaludis)
A25		gpr peptidase (Bacillus megaterium)
A26		omptin (Escherichia coli)
A28	A28A	DNA-damage inducible protein 1 (Saccharomyces cerevisiae)
	A28B	skin SASPase (Mus musculus)
A31		HybD peptidase (Escherichia coli)
A32		PerP peptidase (Caulobacter crescentus)
A36		sporulation factor SpoIIGA (Bacillus subtilis)

Cysteine (C) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
C1	C1A	papain (Carica papaya)
	C1B	bleomycin hydrolase (Saccharomyces cerevisiae)
C2	C2A	calpain-2 (Homo sapiens)
C3	C3A	poliovirus-type picornain 3C (human poliovirus 1)
	C3B	enterovirus picornain 2A (human poliovirus 1)
	C3C	foot-and-mouth disease virus picornain 3C (foot-and-mouth disease virus)
	C3D	cowpea mosaic comovirus-type picornain 3C (cowpea mosaic virus)
	C3E	hepatitis A virus-type picornain 3C (hepatitis A virus)
	C3F	parechovirus picornain 3C (human parechovirus 1)
	C3G	rice tungro spherical virus-type peptidase (rice tungro spherical virus)
	C3H	grapevine fanleaf-type nepovirus picornain 3C (grapevine fanleaf virus)
C4		nuclear-inclusion-a peptidase (plum pox virus)
C5		adenain (human adenovirus type 2)
C6		potato virus Y-type helper component peptidase (potato virus Y)
C7		chestnut blight fungus virus p29 peptidase (Cryphonectria hypovirus)
C8		chestnut blight fungus virus p48 peptidase (Cryphonectria hypovirus 1)
C9		sindbis virus-type nsP2 peptidase (Sindbis virus)
C10		streptopain (Streptococcus pyogenes)
C11		clostripain (Clostridium histolyticum)
C12		ubiquitinyl hydrolase-L1 (Homo sapiens)
C13		legumain (Canavalia ensiformis)
C14	C14A	caspase-1 (Rattus norvegicus)
	C14B	metacaspase Yca1 (Saccharomyces cerevisiae)
C15		pyroglutamyl-peptidase I (Bacillus amyloliquefaciens)
C16	C16A	murine hepatitis coronavirus papain-like peptidase 1 (murine hepatitis virus)
	C16B	murine hepatitis coronavirus papain-like peptidase 2 (murine hepatitis virus)
C18		hepatitis C virus peptidase 2 (hepatitis C virus)
C19		ubiquitin-specific peptidase 14 (Homo sapiens)
C21		tymovirus peptidase (turnip yellow mosaic virus)
C23		carlavirus peptidase (apple stem pitting virus)
C24		rabbit hemorrhagic disease virus 3C-like peptidase (rabbit hemorrhagic disease virus)
C25		gingipain RgpA (Porphyromonas gingivalis)
C26		gamma-glutamyl hydrolase (Rattus norvegicus)
C27		rubella virus peptidase (Rubella virus)
C28		foot-and-mouth disease virus L-peptidase (foot-and-mouth disease virus)
C30		porcine transmissible gastroenteritis virus-type main peptidase (transmissible gastroenteritis virus)
C31		porcine reproductive and respiratory syndrome arterivirus-type cysteine peptidase alpha (lactate-dehydrogenase-elevating virus)
C32		equine arteritis virus-type cysteine peptidase (porcine reproductive and respiratory syndrome virus)
C33		equine arteritis virus Nsp2-type cysteine peptidase (equine arteritis virus)
C36		beet necrotic yellow vein furovirus-type papain-like peptidase (beet necrotic yellow vein virus)
C37		calicivirin (Southampton virus)
C39		bacteriocin-processing peptidase (Pediococcus acidilactici)
C40		dipeptidyl-peptidase VI (Lysinibacillus sphaericus)
C42		beet yellows virus-type papain-like peptidase (beet yellows virus)
C44		amidophosphoribosyltransferase precursor (Homo sapiens)
C45		acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase precursor (Penicillium chrysogenum)
C46		hedgehog protein (Drosophila melanogaster)
C47		staphopain A (Staphylococcus aureus)
C48		Ulp1 peptidase (Saccharomyces cerevisiae)
C50		separase (Saccharomyces cerevisiae)
C51		D-alanyl-glycyl peptidase (Staphylococcus aureus)
C53		pestivirus Npro peptidase (classical swine fever virus)
C54		autophagin-1 (Homo sapiens)
C55		YopJ protein (Yersinia pseudotuberculosis)
C56		PfpI peptidase (Pyrococcus furiosus)
C57		vaccinia virus I7L processing peptidase (Vaccinia virus)
C58	C58A	YopT peptidase (Yersinia pestis)
	C58B	HopN1 peptidase (Pseudomonas syringae)
C59		penicillin V acylase precursor (Lysinibacillus sphaericus)
C60	C60A	sortase A (Staphylococcus aureus)
	C60B	sortase B (Staphylococcus aureus)
C62		gill-associated virus 3C-like peptidase (gill-associated virus)
C63		African swine fever virus processing peptidase (African swine fever virus)
C64		Cezanne peptidase (Homo sapiens)
C65		otubain-1 (Homo sapiens)
C66		IdeS peptidase (Streptococcus pyogenes)
C67		CylD peptidase (Homo sapiens)
C69		dipeptidase A (Lactobacillus helveticus)

C70		AvrRpt2 peptidase (Pseudomonas syringae)
C71		pseudomurein endoisopeptidase Pei (Methanobacterium phage psiM2)
C74		pestivirus NS2 peptidase (bovine viral diarrhea virus 1)
C75		AgrB peptidase (Staphylococcus aureus)
C76		viral tegument protein deubiquitinylating peptidase (human herpesvirus 1)
C78		UfSP1 peptidase (Mus musculus)
C79		ElaD peptidase (Escherichia coli)
C80		RTX self-cleaving toxin (Vibrio cholerae)
C82	C82A	L,D-transpeptidase (Enterococcus faecium)
C83		gamma-glutamylcysteine dipeptidyltranspeptidase (Nostoc sp. PCC 7120)
C84		prtH peptidase (Tannerella forsythia)
C85	C85A	OTLD1 deubiquitinylating enzyme (Homo sapiens)
	C85B	OTU1 peptidase (Saccharomyces cerevisiae)
C86		ataxin-3 (Homo sapiens)
C87		nairovirus deubiquitinylating peptidase (Crimean-Congo hemorrhagic fever virus)
C89		acid ceramidase precursor (Homo sapiens)
C93		LapG peptidase (Pseudomonas fluorescens)
C95		lysosomal 66.3 kDa protein (Mus musculus)
C96		McjB peptidase (Escherichia coli)
C97		DeSI-1 peptidase (Mus musculus)
C98		USPL1 peptidase (Homo sapiens)
C99		iflavirus processing peptidase (Ectropis obliqua picorna-like virus)
C101		OTULIN peptidase (Homo sapiens)
C102		GtgE peptidase (Salmonella enterica)
C104		PlyC phage lysin (Streptococcus phage C1)
C105		papain-like peptidase 1 alpha (simian hemorrhagic fever virus)
C107		alphamesonivirus 3C-like peptidase (Cavally virus)
C108		Prp peptidase (Staphylococcus aureus)
C110		kyphoscoliosis peptidase (Mus musculus)
C111		coagulation factor XIIIa (Homo sapiens)
C113		IgdE peptidase (Streptococcus suis)

Glutamic (G) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
G1		scytalidoglutamic peptidase (Scytalidium lignicolum)
G2		pre-neck appendage protein (bacteriophage phi-29)

Metallo (M) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
M1		aminopeptidase N (Homo sapiens)
M2		angiotensin-converting enzyme peptidase unit 1 (Homo sapiens)
M3	M3A	thimet oligopeptidase (Rattus norvegicus)
	M3B	oligopeptidase F (Lactococcus lactis)
M4		thermolysin (Bacillus thermoproteolyticus)
M5		mycolysin (Streptomyces cacaoi)
M6		immune inhibitor A peptidase (Bacillus thuringiensis)
M7		snapalysin (Streptomyces lividans)
M8		leishmanolysin (Leishmania major)
M9	M9A	bacterial collagenase V (Vibrio alginolyticus)
	M9B	bacterial collagenase H (Clostridium histolyticum)
M10	M10A	matrix metallopeptidase-1 (Homo sapiens)
	M10B	serralysin (Serratia marcescens)
	M10C	fragilysin (Bacteroides fragilis)
M11		gametolysin (Chlamydomonas reinhardtii)
M12	M12A	astacin (Astacus astacus)
	M12B	adamalysin (Crotalus adamanteus)
M13		neprilysin (Homo sapiens)
M14	M14A	carboxypeptidase A1 (Homo sapiens)
	M14B	carboxypeptidase E (Bos taurus)
	M14C	gamma-D-glutamyl--meso-diaminopimelate peptidase I (Lysinibacillus sphaericus)
	M14D	cytosolic carboxypeptidase 6 (Homo sapiens)
M15	M15A	zinc D-Ala-D-Ala carboxypeptidase (Streptomyces albus)
	M15B	vanY D-Ala-D-Ala carboxypeptidase (Enterococcus faecium)
	M15C	Ply118 L-Ala-D-Glu peptidase (bacteriophage A118)
	M15D	vanX D-Ala-D-Ala dipeptidase (Enterococcus faecium)
M16	M16A	pitrilysin (Escherichia coli)
	M16B	mitochondrial processing peptidase beta-subunit (Saccharomyces cerevisiae)
	M16C	eupitrilysin (Homo sapiens)
M17		leucine aminopeptidase 3 (Bos taurus)
M18		aminopeptidase I (Saccharomyces cerevisiae)
M19		membrane dipeptidase (Homo sapiens)
M20	M20A	glutamate carboxypeptidase (Pseudomonas sp.)
	M20B	peptidase T (Escherichia coli)
	M20C	Xaa-His dipeptidase (Escherichia coli)
	M20D	carboxypeptidase Ss1 (Sulfolobus solfataricus)
	M20F	carnosine dipeptidase II (Mus musculus)
M23	M23A	beta-lytic metallopeptidase (Achromobacter lyticus)
	M23B	lysostaphin (Staphylococcus simulans)
M24	M24A	methionyl aminopeptidase 1 (Escherichia coli)
	M24B	aminopeptidase P (Escherichia coli)
M26		IgA1-specific metallopeptidase (Streptococcus sanguinis)
M27		tentoxilysin (Clostridium tetani)
M28	M28A	aminopeptidase S (Streptomyces griseus)
	M28B	glutamate carboxypeptidase II (Homo sapiens)
	M28C	IAP aminopeptidase (Escherichia coli)
	M28D	aminopeptidase ES-62 (Acanthocheilonema viteae)
	M28E	aminopeptidase Ap1 (Vibrio proteolyticus)
	M28F	ywaD peptidase (Bacillus subtilis)
M29		aminopeptidase T (Thermus aquaticus)
M30		hyicolysin (Staphylococcus hyicus)
M32		carboxypeptidase Taq (Thermus aquaticus)
M34		anthrax lethal factor (Bacillus anthracis)
M35		deuterolysin (Aspergillus flavus)
M36		fungalysin (Aspergillus fumigatus)
M38		isoaspartyl dipeptidase (Escherichia coli)
M41		FtsH peptidase (Escherichia coli)
M42		glutamyl aminopeptidase (Lactococcus lactis)
M43	M43A	cytophagalysin (Cytophaga sp.)

	M43B	pappalysin-1 (Homo sapiens)
M44		pox virus metallopeptidase (Vaccinia virus)
M48	M48A	Ste24 peptidase (Saccharomyces cerevisiae)
	M48B	HtpX peptidase (Escherichia coli)
	M48C	Oma1 peptidase (Saccharomyces cerevisiae)
M49		dipeptidyl-peptidase III (Rattus norvegicus)
M50	M50A	site 2 peptidase (Homo sapiens)
	M50B	sporulation factor SpoIVFB (Bacillus subtilis)
M54		archaelysin (Methanocaldococcus jannaschii)
M55		D-aminopeptidase DppA (Bacillus subtilis)
M56		BlaR1 peptidase (Staphylococcus aureus)
M57		prtB g.p. (Myxococcus xanthus)
M60		enhancin (Lymantria dispar nucleopolyhedrovirus)
M61		glycyl aminopeptidase (Sphingomonas capsulata)
M64		IgA peptidase (Clostridium ramosum)
M66		StcE peptidase (Escherichia coli)
M67	M67A	RPN11 peptidase (Saccharomyces cerevisiae)
	M67B	JAMM-like protein (Archaeoglobus fulgidus)
	M67C	STAMBP isopeptidase (Homo sapiens)
M72		peptidyl-Asp metallopeptidase (Pseudomonas aeruginosa)
M73		camelysin (Bacillus cereus)
M74		murein endopeptidase (Escherichia coli)
M75		imelysin (Pseudomonas aeruginosa)
M76		Atp23 peptidase (Homo sapiens)
M77		tryptophanyl aminopeptidase 7-DMATS-type peptidase (Aspergillus fumigatus)
M78		ImmA peptidase (Bacillus subtilis)
M79		RCE1 peptidase (Saccharomyces cerevisiae)
M80		Wss1 peptidase (Saccharomyces cerevisiae)
M81		microcystinase MlrC (Sphingomonas sp. ACM-3962)
M82		PrsW peptidase (Bacillus subtilis)
M84		MpriBi peptidase (Bacillus intermedius)
M85		NleC peptidase (Escherichia coli)
M86		PghP gamma-polyglutamate hydrolase (Bacillus phage phiNIT1)
M87		chloride channel accessory protein 1 (Homo sapiens)
M88		IMPa peptidase (Pseudomonas aeruginosa)
M90		MtfA peptidase (Escherichia coli)
M91		NleD peptidase (Escherichia coli)
M93		BACCAC_01431 g.p. and similar (Bacteroides caccae)
M95		selecase (Methanocaldococcus jannaschii)
M96		Tiki1 peptidase (Homo sapiens)
M97		EcxAB peptidase (Escherichia coli)
M98		YghJ g.p. (Escherichia coli)
M99		Csd4 peptidase (Helicobacter pylori)

Asparagine (N) Peptide Lyases

FAMILY	SUBFAMILY	TYPE ENZYME
N1		nodavirus peptide lyase (flock house virus)
N2		tetravirus coat protein (Nudaurelia capensis omega virus)
N4		Tsh-associated self-cleaving domain and similar (Escherichia coli)
N5		picobirnavirus self-cleaving protein (Human picobirnavirus)
N6		YscU protein (Yersinia pseudotuberculosis)
N7		reovirus type 1 coat protein (Mammalian orthoreovirus 1)
N8		poliovirus capsid VP0-type self-cleaving protein (human poliovirus 1)
N9		intein-containing V-type proton ATPase catalytic subunit A (Saccharomyces cerevisiae)
N10		intein-containing replicative DNA helicase precursor (Synechocystis sp. PCC 6803)
N11		intein-containing chloroplast ATP-dependent peptide lyase (Chlamydomonas eugametos)

Mixed (P) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
P1		DmpA aminopeptidase (Ochrobactrum anthropi)
P2	P2A	EGF-like module containing mucin-like hormone receptor-like 2 (Homo sapiens)
	P2B	polycystin-1 (Homo sapiens)

Serine (S) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
S1	S1A	chymotrypsin A (Bos taurus)
	S1B	glutamyl endopeptidase I (Staphylococcus aureus)
	S1C	DegP peptidase (Escherichia coli)
	S1D	lysyl endopeptidase (Achromobacter lyticus)
	S1E	streptogrisin A (Streptomyces griseus)
	S1F	astrovirus serine peptidase (Mamastrovirus 1)
S3		togavirin (Sindbis virus)
S6		IgA1-specific serine peptidase (Neisseria gonorrhoeae)
S7		flavivirin (yellow fever virus)
S8	S8A	subtilisin Carlsberg (Bacillus licheniformis)
	S8B	kexin (Saccharomyces cerevisiae)
S9	S9A	prolyl oligopeptidase (Sus scrofa)
	S9B	dipeptidyl-peptidase IV (Homo sapiens)
	S9C	acylaminoacyl-peptidase (Homo sapiens)
	S9D	glutamyl endopeptidase C (Arabidopsis thaliana)
S10		carboxypeptidase Y (Saccharomyces cerevisiae)
S11		D-Ala-D-Ala carboxypeptidase A (Geobacillus stearothermophilus)
S12		D-Ala-D-Ala carboxypeptidase B (Streptomyces lividans)
S13		D-Ala-D-Ala peptidase C (Escherichia coli)
S14		peptidase Clp (Escherichia coli)
S15		Xaa-Pro dipeptidyl-peptidase (Lactococcus lactis)
S16		Lon-A peptidase (Escherichia coli)
S21		cytomegalovirus assemblin (human herpesvirus 5)
S24		repressor LexA (Escherichia coli)
S26	S26A	signal peptidase I (Escherichia coli)
	S26B	signalase 21 kDa component (Saccharomyces cerevisiae)
	S26C	TraF peptidase (Escherichia coli)
S28		lysosomal Pro-Xaa carboxypeptidase (Homo sapiens)
S29		hepacivirin (hepatitis C virus)
S30		potyvirus P1 peptidase (plum pox virus)
S31		pestivirus NS3 polyprotein peptidase (bovine viral diarrhea virus 1)
S32		equine arteritis virus serine peptidase (equine arteritis virus)
S33		prolyl aminopeptidase (Neisseria gonorrhoeae)
S37		PS-10 peptidase (Streptomyces lividans)

S39	S39A	sobemovirus peptidase (cocksfoot mottle virus)
	S39B	luteovirus peptidase (potato leaf roll luteovirus)
S41	S41A	C-terminal processing peptidase-1 (Escherichia coli)
	S41B	tricorn core peptidase (Thermoplasma acidophilum)
S45		penicillin G acylase precursor (Escherichia coli)
S46		dipeptidyl-peptidase 7 (Porphyromonas gingivalis)
S48		HetR putative peptidase (Anabaena variabilis)
S49	S49A	signal peptide peptidase A (Escherichia coli)
	S49B	protein C (bacteriophage lambda)
	S49C	archaean signal peptide peptidase 1 (Pyrococcus horikoshii)
S50		infectious pancreatic necrosis birnavirus Vp4 peptidase (infectious pancreatic necrosis virus)
S51		dipeptidase E (Escherichia coli)
S53		sedolisin (Pseudomonas sp. 101)
S54		rhomboid-1 (Drosophila melanogaster)
S55		SpoIVB peptidase (Bacillus subtilis)
S59		nucleoporin 145 (Homo sapiens)
S60		lactoferrin (Homo sapiens)
S62		influenza A PA peptidase (influenza A virus)
S64		Ssy5 peptidase (Saccharomyces cerevisiae)
S65		picornain-like cysteine peptidase (Breda virus)
S66		murein tetrapeptidase LD-carboxypeptidase (Pseudomonas aeruginosa)
S68		PIDD auto-processing protein unit 1 (Homo sapiens)
S69		Tellina virus 1 VP4 peptidase (Tellina virus 1)
S71		MUC1 self-cleaving mucin (Homo sapiens)
S72		dystroglycan (Homo sapiens)
S73		gpO peptidase (Enterobacteria phage P2)
S74		Escherichia coli phage K1F endosialidase CIMCD self-cleaving protein (Enterobacteria phage K1F)
S75		White bream virus serine peptidase (White bream virus)
S77		prohead peptidase gp21 (Enterobacteria phage T4)
S78		prohead peptidase (Enterobacteria phage HK97)
S79		CARD8 self-cleaving protein (Homo sapiens)
S80		prohead peptidase gp175 (Pseudomonas phage phiKZ)
S81		destabilase (Hirudo medicinalis)

Threonine (T) Peptidases

FAMILY	SUBFAMILY	TYPE ENZYME
T1	T1A	archaean proteasome, beta component (Thermoplasma acidophilum)
	T1B	HslV component of HslUV peptidase (Escherichia coli)
T2		glycosylasparaginase precursor (Homo sapiens)
T3		gamma-glutamyltransferase 1 (Escherichia coli)
T5		ornithine acetyltransferase precursor (Saccharomyces cerevisiae)
T7		CwpV self-cleaving threonine peptidase (Peptoclostridium difficile)

Peptidases of Unknown Catalytic Type

FAMILY	SUBFAMILY	TYPE ENZYME
U32		collagenase (Porphyromonas gingivalis)
U40		protein P5 murein endopeptidase (bacteriophage phi-6)
U49		Lit peptidase (Escherichia coli)
U56		homomultimeric peptidase (Thermotoga maritima)
U57		yabG protein (Bacillus subtilis)
U62		microcin-processing peptidase 1 (Escherichia coli)
U69		AIDA-I self-cleaving autotransporter protein (Escherichia coli)
U72		Dop isopeptidase (Mycobacterium tuberculosis)
U73		small protease (Pseudomonas aeruginosa)

Table S7

strain	replicon	CRISPR type	CRISPR sub-type	repeat number	repeat length	repeat sequence	avg. length of spacers	cas 1	cas 3	cas 9	cas 10
778	chromosome	II	II-A	7	34	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGA	32	Y		Y	
866	chromosome	II	II-A	59	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	30	Y		Y	
866	chromosome	III	III-A	19	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
01M14315	chromosome	II	II-A	12	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
01M14315	chromosome	III	III-A	13	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	39	Y			Y
ACS116	chromosome	II	II-A	17	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	29	Y		Y	
ACS116	chromosome	III	III-A	9	35	TTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			
AH4231	chromosome	II	II-A	60	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	30	Y		Y	
AH4231	chromosome	III	III-A	26	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
AH4331	chromosome	II	II-A	60	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	30	Y		Y	
AH4331	chromosome	III	III-A	26	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
AH43324	chromosome	II	II-A	28	36	GTTTCAGAAGTATGTTAAATCAATAAGGTTAAGACC	30	Y		Y	
AH43348	chromosome	II	II-A	62	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	30	Y		Y	
AH43348	chromosome	III	III-A	31	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
ATCC11741	chromosome	III	III-A	12	36	GTTTTCGTCTCCTTCATTCGGAGATATGTTCTTATT	37	Y			Y
CCuG2753OB	chromosome	II	II-A	19	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
CCuG38008	chromosome	II	II-A	29	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
CCuG44481	NA	NA	NA	0	0	NA	0				
CCuG45735	chromosome	II	II-A	40	36	GTTTCAGAAGGATGTTAAATCAATTAGGTTAAGACT	30	Y		Y	
CCuG45735	chromosome	III	III-A	19	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
CCuG47171	chromosome	III	III-A	11	36	GTTTTCGTCTCCTTTATTCGGAGATATGTTCTTACT	37	Y			Y
CCuG47825	chromosome	II	II-A	23	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
CCuG47825	chromosome	undefined	undefined	12	37	AGTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	40				
CCuG47826	chromosome	II	II-A	22	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
CCuG47826	chromosome	III	III-A	19	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	37	Y			Y
CECT5713	chromosome	II	II-A	28	36	GTTTCAGAAGTATGTTAAATCAATAAGGTTAAGACC	30	Y		Y	
cp400	chromosome	III	III-A	28	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
DSM18933	NA	NA	NA	0	0	NA	0				
DSM20492	chromosome	II	II-A	17	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
DSM20492	chromosome	III	III-A	22	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
DSM20554	chromosome	III	III-A	22	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	37	Y			Y
DSM20555	chromosome	III	III-A	12	36	GTTTTCGTCTCCTTCATTCGGAGATATGTTCTTATT	37	Y			Y
GJ24	NA	NA	NA	0	0	NA	0				
gul1	chromosome	III	III-A	12	36	GTTTTCGTCTCCTTCATTCGGAGATATGTTCTTATT	37	Y			Y
gul2	chromosome	III	III-A	12	36	GTTTTCGTCTCCTTCATTCGGAGATATGTTCTTATT	37	Y			Y
JCM1040	chromosome	II	II-A	15	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	29	Y		Y	
JCM1040	chromosome	III	III-A	18	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
JCM1042	chromosome	II	II-A	19	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
JCM1044	chromosome	II	II-A	19	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
JCM1045	NA	NA	NA	0	0	NA	0				
JCM1046	chromosome	III	III-A	27	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	38	Y			Y
JCM1047	NA	NA	NA	0	0	NA	0				
JCM1230	chromosome	II	II-A	40	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGGCC	29	Y		Y	
L21	chromosome	II	II-A	26	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	29	Y		Y	
L21	chromosome	III	III-A	50	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	37	Y			Y
LMG14476	NA	NA	NA	0	0	NA	0				
LMG14477	NA	NA	NA	0	0	NA	0				
NCIMB702343	chromosome	II	II-A	41	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
NCIMB702343	chromosome	undescribed	undescribed	11	24	GTTTCAGAAGTATGTTAAATCAAT	41				
NCIMB702343	chromosome	III	III-A	20	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	37	Y			Y
NCIMB702343	chromosome	undescribed	undescribed	6	35	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGAC	31				
NCIMB8816	chromosome	II	II-A	8	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
NCIMB8817	chromosome	II	II-A	50	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	30	Y		Y	
NCIMB8818	chromosome	II	II-A	8	36	GTTTCAGAAGTATGTTAAATCAATAAGGTTAAGACC	30	Y		Y	
NIAS840	chromosome	undefined	undefined	18	31	TCAAGTTCCTTAAGTGAAAGCTTGAGTACAT	40				
NIAS840	chromosome	III	III-A	9	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	36	Y			Y
NIAS840	megaplasmid	undefined	undefined	5	37	GTCTTACACCTATGTCAATTCAACTAGGTTCAGAACC	29				
Ren	chromosome	II	II-A	11	36	GTTTCAGAAGTATGTTAAATCAATTAGGTTAAGACC	29	Y			Y
Ren	chromosome	III	III-A	18	36	GTTTTCGTCTCCTATATTCGGAGATATGTTCTTACT	37				
SMXD51	chromosome	II	II-A	25	36	GTTTCAGAAGTATGTTAAATCAATAAGGTTAAGACC	29	Y		Y	
UCC118	chromosome	II	II-A	28	36	GTTTCAGAAGTATGTTAAATCAATAAGGTTAAGACC	30	Y		Y	

Table S8

query gene	% identity	query coverage	e-value	query gene length	functional annotation
01M14315_ORF_49	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
01M14315_ORF_69	49	498	2.00E-170	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
01M14315_ORF_212	48	233	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
01M14315_ORF_611	46	418	8.00E-123	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
01M14315_ORF_653	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
01M14315_ORF_936	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
01M14315_ORF_940	41	569	4.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
01M14315_ORF_1128	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
01M14315_ORF_1278	62	61	2.00E-22	73	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
01M14315_ORF_1354	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
01M14315_ORF_1416	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
01M14315_ORF_1628	43	219	2.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
778_ORF_53	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
778_ORF_72	50	500	2.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
778_ORF_226	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
778_ORF_512	46	418	1.00E-120	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
778_ORF_610	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
778_ORF_828	40	573	2.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
778_ORF_832	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
778_ORF_1079	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
778_ORF_1272	40	222	8.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
778_ORF_1335	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
778_ORF_1549	43	221	6.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
866_ORF_51	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
866_ORF_106	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
866_ORF_107	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
866_ORF_271	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
866_ORF_642	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
866_ORF_682	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
866_ORF_913	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
866_ORF_917	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
866_ORF_1176	41	233	6.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
866_ORF_1362	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
866_ORF_1373	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
866_ORF_1455	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
866_ORF_1663	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ACS116_ORF_52	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
ACS116_ORF_71	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
ACS116_ORF_72	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
ACS116_ORF_228	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
ACS116_ORF_651	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
ACS116_ORF_722	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
ACS116_ORF_1019	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
ACS116_ORF_1023	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
ACS116_ORF_1271	40	187	3.00E-41	191	vanRA, also known as vanR, is a vanR variant found in the vanA gene cluster
ACS116_ORF_1293	40	223	2.00E-49	223	vanRE is a vanR variant found in the vanE gene cluster
ACS116_ORF_1328	45	286	1.00E-80	299	MprF is a integral membrane protein that modifies the negatively-charged phosphatidylglycerol on the membrane surface
ACS116_ORF_1482	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ACS116_ORF_1493	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
ACS116_ORF_1575	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ACS116_ORF_1792	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4231_ORF_45	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
AH4231_ORF_65	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH4231_ORF_66	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH4231_ORF_231	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
AH4231_ORF_639	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
AH4231_ORF_680	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
AH4231_ORF_1023	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
AH4231_ORF_1027	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
AH4231_ORF_1288	41	233	6.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
AH4231_ORF_1471	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4231_ORF_1482	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
AH4231_ORF_1564	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4231_ORF_1777	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4231_ORF_1862	40	227	1.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
AH4331_ORF_45	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
AH4331_ORF_63	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH4331_ORF_64	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH4331_ORF_230	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
AH4331_ORF_647	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
AH4331_ORF_687	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
AH4331_ORF_994	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
AH4331_ORF_998	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
AH4331_ORF_1263	41	233	6.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
AH4331_ORF_1446	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4331_ORF_1457	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
AH4331_ORF_1539	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4331_ORF_1752	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH4331_ORF_1832	40	227	1.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
AH43324_ORF_52	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
AH43324_ORF_70	50	498	3.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH43324_ORF_231	48	228	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
AH43324_ORF_630	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
AH43324_ORF_670	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
AH43324_ORF_969	40	583	1.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
AH43324_ORF_973	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
AH43324_ORF_1227	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
AH43324_ORF_1280	40	400	2.00E-90	415	MprF is a integral membrane protein that modifies the negatively-charged phosphatidylglycerol on the membrane surface
AH43324_ORF_1428	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH43324_ORF_1439	45	305	1.00E-79	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
AH43324_ORF_1522	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH43324_ORF_1765	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH43348_ORF_45	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
AH43348_ORF_64	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH43348_ORF_65	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
AH43348_ORF_232	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
AH43348_ORF_640	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
AH43348_ORF_680	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
AH43348_ORF_992	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
AH43348_ORF_996	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
AH43348_ORF_1259	41	233	6.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
AH43348_ORF_1447	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH43348_ORF_1458	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
AH43348_ORF_1539	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
AH43348_ORF_1755	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ATCC11741_ORF_54	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
ATCC11741_ORF_71	50	498	3.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
ATCC11741_ORF_216	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster

ATCC11741_ORF_559	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
ATCC11741_ORF_600	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
ATCC11741_ORF_824	41	573	9.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
ATCC11741_ORF_828	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
ATCC11741_ORF_1079	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
ATCC11741_ORF_1264	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ATCC11741_ORF_1327	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
ATCC11741_ORF_1544	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG2753OB_ORF_53	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG2753OB_ORF_73	49	500	1.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG2753OB_ORF_328	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG2753OB_ORF_506	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG2753OB_ORF_546	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG2753OB_ORF_777	40	583	4.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG2753OB_ORF_782	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG2753OB_ORF_1031	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG2753OB_ORF_1212	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG2753OB_ORF_1275	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG2753OB_ORF_1501	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG38008_ORF_55	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG38008_ORF_70	49	500	2.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG38008_ORF_221	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG38008_ORF_546	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG38008_ORF_586	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG38008_ORF_885	40	583	9.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG38008_ORF_889	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG38008_ORF_1140	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG38008_ORF_1286	60	63	1.00E-22	67	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CCuG38008_ORF_1389	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG38008_ORF_1606	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG38008_ORF_1669	40	227	2.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
CCuG44481_ORF_44	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG44481_ORF_96	50	498	2.00E-172	490	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG44481_ORF_246	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG44481_ORF_661	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG44481_ORF_702	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG44481_ORF_911	40	583	7.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG44481_ORF_915	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG44481_ORF_1148	41	233	1.00E-53	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG44481_ORF_1338	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG44481_ORF_1401	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG44481_ORF_1700	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CCuG44481_ORF_1782	99	458	0	458	TetL is a tetracycline efflux protein found in many species of Gram-negative and Gram-positive bacteria
CCuG44481_ORF_1931	90	641	0	646	TetM is a ribosomal protection protein that confers tetracycline resistance
CCuG44481_ORF_1937	99	215	4.00E-156	215	cat is used to describe many variants of the chloramphenicol acetyltransferase gene in a range of organisms
CCuG44481_ORF_1974	51	306	8.00E-109	326	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
CCuG45735_ORF_51	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG45735_ORF_70	55	95	7.00E-31	99	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG45735_ORF_71	49	391	5.00E-131	387	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG45735_ORF_227	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG45735_ORF_644	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG45735_ORF_685	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG45735_ORF_918	40	583	9.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG45735_ORF_922	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG45735_ORF_1183	41	233	6.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG45735_ORF_1365	40	219	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG45735_ORF_1376	45	302	2.00E-80	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
CCuG45735_ORF_1461	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG45735_ORF_1681	43	219	4.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG45735_ORF_1865	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CCuG47171_ORF_51	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG47171_ORF_73	49	498	2.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG47171_ORF_217	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG47171_ORF_594	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG47171_ORF_634	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG47171_ORF_889	40	583	2.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG47171_ORF_893	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG47171_ORF_1141	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG47171_ORF_1371	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47171_ORF_1434	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47171_ORF_1663	44	204	3.00E-51	219	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47171_ORF_1918	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CCuG47171_ORF_2091	41	231	2.00E-53	229	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_286	41	136	4.00E-26	154	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_287	49	81	1.00E-20	102	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_505	43	90	4.00E-21	94	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_506	47	122	3.00E-30	128	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_570	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47825_ORF_773	46	418	1.00E-122	438	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG47825_ORF_817	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG47825_ORF_1076	41	493	7.00E-129	494	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG47825_ORF_1080	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG47825_ORF_1351	45	225	1.00E-62	248	vanRM is a vanR variant found in the vanM gene cluster
CCuG47825_ORF_1369	50	498	7.00E-174	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG47825_ORF_1449	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG47825_ORF_1451	44	77	0.00E+00	93	vanSL is a vanS variant found in the vanL gene cluster
CCuG47825_ORF_1688	40	164	3.00E-32	172	A type III ABC transporter, identified on the novobiocin biosynthetic gene cluster
CCuG47825_ORF_1696	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG47825_ORF_1830	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CCuG47826_ORF_51	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CCuG47826_ORF_68	50	498	7.00E-174	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CCuG47826_ORF_234	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CCuG47826_ORF_242	40	164	3.00E-32	172	A type III ABC transporter, identified on the novobiocin biosynthetic gene cluster
CCuG47826_ORF_556	46	418	8.00E-123	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CCuG47826_ORF_596	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CCuG47826_ORF_907	41	569	3.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CCuG47826_ORF_911	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CCuG47826_ORF_1164	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CCuG47826_ORF_1452	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47826_ORF_1514	43	222	3.00E-60	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47826_ORF_1724	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CCuG47826_ORF_1842	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
CECT5713_ORF_44	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
CECT5713_ORF_64	50	500	6.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
CECT5713_ORF_249	48	231	4.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
CECT5713_ORF_699	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
CECT5713_ORF_742	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
CECT5713_ORF_1052	41	127	3.00E-28	136	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
CECT5713_ORF_1053	45	350	3.00E-102	412	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex

CECT5713_ORF_1057	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
CECT5713_ORF_1328	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
CECT5713_ORF_1549	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CECT5713_ORF_1563	45	305	1.00E-79	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
CECT5713_ORF_1647	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
CECT5713_ORF_1903	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
cp400_ORF_61	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
cp400_ORF_244	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
cp400_ORF_747	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
cp400_ORF_789	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
cp400_ORF_799	40	583	2.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
cp400_ORF_803	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
cp400_ORF_1294	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
cp400_ORF_1551	40	222	1.00E-45	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
cp400_ORF_1615	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
cp400_ORF_2079	60	60	2.00E-20	71	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
cp400_ORF_2106	50	498	8.00E-168	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
DSM18933_ORF_149	42	107	3.00E-23	110	EmrE is a small multidrug transporter that functions as a homodimer
DSM18933_ORF_286	47	226	1.00E-67	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM18933_ORF_502	44	225	1.00E-59	236	vanRF is a vanR variant found in the vanF gene cluster
DSM18933_ORF_616	49	228	3.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
DSM18933_ORF_1243	44	417	7.00E-115	422	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
DSM18933_ORF_1422	46	645	0	766	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
DSM18933_ORF_1449	68	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
DSM20492_ORF_47	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
DSM20492_ORF_69	50	500	5.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
DSM20492_ORF_218	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
DSM20492_ORF_577	46	418	3.00E-121	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
DSM20492_ORF_617	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
DSM20492_ORF_862	40	583	9.00E-148	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
DSM20492_ORF_866	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
DSM20492_ORF_1098	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
DSM20492_ORF_1287	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20492_ORF_1350	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20492_ORF_1554	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20554_ORF_46	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
DSM20554_ORF_68	50	500	5.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
DSM20554_ORF_208	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
DSM20554_ORF_565	46	418	3.00E-121	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
DSM20554_ORF_681	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
DSM20554_ORF_910	40	583	9.00E-148	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
DSM20554_ORF_914	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
DSM20554_ORF_1106	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
DSM20554_ORF_1298	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20554_ORF_1361	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20554_ORF_1581	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20555_ORF_55	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
DSM20555_ORF_72	50	498	3.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
DSM20555_ORF_212	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
DSM20555_ORF_554	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
DSM20555_ORF_594	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
DSM20555_ORF_815	41	573	9.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
DSM20555_ORF_819	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
DSM20555_ORF_1069	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
DSM20555_ORF_1257	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20555_ORF_1319	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
DSM20555_ORF_1535	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
GJ24_ORF_21	50	498	2.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
GJ24_ORF_180	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
GJ24_ORF_620	46	418	1.00E-120	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
GJ24_ORF_661	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
GJ24_ORF_870	40	583	6.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
GJ24_ORF_874	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
GJ24_ORF_1110	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
GJ24_ORF_1344	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
GJ24_ORF_1407	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
GJ24_ORF_1736	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
gul1_ORF_54	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
gul1_ORF_71	50	498	3.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
gul1_ORF_215	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
gul1_ORF_519	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
gul1_ORF_560	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
gul1_ORF_783	41	573	9.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
gul1_ORF_787	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
gul1_ORF_1039	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
gul1_ORF_1226	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
gul1_ORF_1288	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
gul1_ORF_1505	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
gul2_ORF_55	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
gul2_ORF_72	50	498	3.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
gul2_ORF_217	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
gul2_ORF_559	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
gul2_ORF_600	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
gul2_ORF_825	41	573	9.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
gul2_ORF_829	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
gul2_ORF_1081	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
gul2_ORF_1268	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
gul2_ORF_1330	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
gul2_ORF_1547	43	221	1.00E-56	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1040_ORF_49	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1040_ORF_70	50	498	4.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1040_ORF_215	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1040_ORF_599	46	418	8.00E-123	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1040_ORF_639	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1040_ORF_876	41	569	4.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1040_ORF_880	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1040_ORF_1123	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1040_ORF_1300	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1040_ORF_1362	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1040_ORF_1589	43	219	2.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1040_ORF_1656	40	227	1.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
JCM1042_ORF_53	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1042_ORF_73	49	500	1.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1042_ORF_218	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1042_ORF_549	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1042_ORF_590	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1042_ORF_819	40	583	4.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1042_ORF_824	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1042_ORF_1071	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster

JCM1042_ORF_1252	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1042_ORF_1315	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1042_ORF_1545	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1044_ORF_54	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1044_ORF_71	49	500	1.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1044_ORF_216	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1044_ORF_504	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1044_ORF_544	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1044_ORF_778	40	583	4.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1044_ORF_783	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1044_ORF_1031	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1044_ORF_1210	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1044_ORF_1273	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1044_ORF_1498	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1045_ORF_56	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1045_ORF_74	50	498	3.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1045_ORF_215	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1045_ORF_544	46	418	4.00E-121	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1045_ORF_649	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1045_ORF_886	41	583	2.00E-148	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1045_ORF_890	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1045_ORF_1145	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1045_ORF_1343	40	222	5.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1045_ORF_1411	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1046_ORF_54	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1046_ORF_75	50	498	8.00E-168	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1046_ORF_220	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1046_ORF_569	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1046_ORF_609	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1046_ORF_874	40	583	2.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1046_ORF_878	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1046_ORF_1099	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1046_ORF_1251	62	61	2.00E-22	73	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
JCM1046_ORF_1320	40	222	1.00E-45	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1046_ORF_1382	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1046_ORF_1829	90	639	0	644	TetM is a ribosomal protection protein that confers tetracycline resistance
JCM1047_ORF_54	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1047_ORF_91	49	496	8.00E-172	492	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1047_ORF_241	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1047_ORF_664	46	418	7.00E-122	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1047_ORF_720	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1047_ORF_1061	40	579	2.00E-144	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1047_ORF_1065	45	648	0	766	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1047_ORF_1301	41	233	1.00E-53	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1047_ORF_1485	40	222	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1047_ORF_1547	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1047_ORF_2131	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
JCM1047_ORF_2228	90	641	0	646	TetM is a ribosomal protection protein that confers tetracycline resistance
JCM1230_ORF_72	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
JCM1230_ORF_93	50	498	4.00E-172	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
JCM1230_ORF_258	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
JCM1230_ORF_548	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
JCM1230_ORF_588	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
JCM1230_ORF_798	40	583	7.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
JCM1230_ORF_802	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
JCM1230_ORF_1046	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
JCM1230_ORF_1288	40	222	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
JCM1230_ORF_1350	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
L21_ORF_50	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
L21_ORF_67	50	498	7.00E-174	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
L21_ORF_222	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
L21_ORF_688	46	418	8.00E-123	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
L21_ORF_728	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
L21_ORF_969	41	573	4.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
L21_ORF_973	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
L21_ORF_1225	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
L21_ORF_1405	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
L21_ORF_1467	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
L21_ORF_1675	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
L21_ORF_1986	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
LMG14476_ORF_69	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
LMG14476_ORF_90	50	498	8.00E-173	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
LMG14476_ORF_231	48	231	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
LMG14476_ORF_268	40	263	1.00E-59	327	vanHA, also known as vanH, is a vanH variant in the vanA gene cluster
LMG14476_ORF_549	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
LMG14476_ORF_589	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
LMG14476_ORF_968	40	583	7.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
LMG14476_ORF_972	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
LMG14476_ORF_1201	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
LMG14476_ORF_1347	60	61	3.00E-23	68	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
LMG14476_ORF_1417	40	222	6.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
LMG14476_ORF_1480	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
LMG14476_ORF_1711	43	221	4.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
LMG14477_ORF_52	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
LMG14477_ORF_73	50	498	8.00E-173	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
LMG14477_ORF_211	48	231	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
LMG14477_ORF_248	40	263	1.00E-59	327	vanHA, also known as vanH, is a vanH variant in the vanA gene cluster
LMG14477_ORF_543	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
LMG14477_ORF_583	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
LMG14477_ORF_952	40	583	7.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
LMG14477_ORF_956	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
LMG14477_ORF_1185	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
LMG14477_ORF_1330	60	61	3.00E-23	68	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
LMG14477_ORF_1401	40	222	6.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
LMG14477_ORF_1464	44	222	2.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
LMG14477_ORF_1697	43	221	4.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB702343_ORF_53	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
NCIMB702343_ORF_79	48	223	3.00E-63	226	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NCIMB702343_ORF_80	52	183	4.00E-62	183	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NCIMB702343_ORF_236	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
NCIMB702343_ORF_578	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
NCIMB702343_ORF_619	46	418	2.00E-122	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
NCIMB702343_ORF_850	40	583	2.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
NCIMB702343_ORF_854	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
NCIMB702343_ORF_1087	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
NCIMB702343_ORF_1243	62	62	7.00E-23	77	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
NCIMB702343_ORF_1315	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB702343_ORF_1381	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB702343_ORF_1604	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones

NCIMB702343_ORF_1779	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
NCIMB8816_ORF_55	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
NCIMB8816_ORF_76	49	500	1.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NCIMB8816_ORF_221	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
NCIMB8816_ORF_507	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
NCIMB8816_ORF_547	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
NCIMB8816_ORF_782	40	583	4.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
NCIMB8816_ORF_787	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
NCIMB8816_ORF_1032	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
NCIMB8816_ORF_1253	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8816_ORF_1317	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8816_ORF_1545	43	221	2.00E-56	244	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8816_ORF_1605	40	227	1.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
NCIMB8817_ORF_60	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
NCIMB8817_ORF_79	50	498	7.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NCIMB8817_ORF_221	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
NCIMB8817_ORF_610	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
NCIMB8817_ORF_651	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
NCIMB8817_ORF_880	41	579	2.00E-148	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
NCIMB8817_ORF_884	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
NCIMB8817_ORF_1133	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
NCIMB8817_ORF_1309	40	222	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8817_ORF_1371	44	222	9.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8818_ORF_51	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
NCIMB8818_ORF_71	50	498	2.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NCIMB8818_ORF_226	49	231	3.00E-71	229	vanRF is a vanR variant found in the vanF gene cluster
NCIMB8818_ORF_452	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8818_ORF_740	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
NCIMB8818_ORF_780	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
NCIMB8818_ORF_1075	40	583	4.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
NCIMB8818_ORF_1079	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
NCIMB8818_ORF_1330	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
NCIMB8818_ORF_1484	62	61	6.00E-22	73	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
NCIMB8818_ORF_1530	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8818_ORF_1541	45	305	1.00E-79	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
NCIMB8818_ORF_1650	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NCIMB8818_ORF_1908	45	77	0.00E+00	76	A type III ABC transporter, identified on the novobiocin biosynthetic gene cluster
NIAS840_ORF_88	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
NIAS840_ORF_241	49	498	3.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
NIAS840_ORF_403	40	222	1.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NIAS840_ORF_465	44	198	1.00E-52	207	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
NIAS840_ORF_467	98	164	9.00E-119	164	lnuC is a transposon-mediated nucleotidyltransferase found in Streptococcus agalactiae
NIAS840_ORF_661	99	216	1.00E-158	216	vatH is a plasmid-mediated acetyltransferase found in Enterococcus faecium
NIAS840_ORF_662	99	525	0	525	vgaD is an efflux protein expressed in Enterococcus faecium that confers resistance to streptogramin A antibiotics
NIAS840_ORF_793	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
NIAS840_ORF_1075	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
NIAS840_ORF_1168	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
NIAS840_ORF_1378	41	583	1.00E-148	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
NIAS840_ORF_1382	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
NIAS840_ORF_1588	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
Ren_ORF_41	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
Ren_ORF_63	50	498	4.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
Ren_ORF_230	48	233	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
Ren_ORF_651	46	418	8.00E-123	440	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
Ren_ORF_691	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
Ren_ORF_929	41	569	4.00E-147	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
Ren_ORF_933	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
Ren_ORF_1188	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
Ren_ORF_1386	40	222	3.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
Ren_ORF_1448	44	222	4.00E-62	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
Ren_ORF_1680	49	81	7.00E-21	102	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
Ren_ORF_1681	46	110	1.00E-26	131	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
Ren_ORF_1761	58	67	8.00E-24	84	VanU is a transcriptional activator of the vanG operon of vancomycin resistance genes
Ren_ORF_1999	40	227	1.00E-54	233	vanRM is a vanR variant found in the vanM gene cluster
SMXD51_ORF_100	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
SMXD51_ORF_205	50	498	1.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
SMXD51_ORF_366	48	233	1.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
SMXD51_ORF_691	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
SMXD51_ORF_758	40	222	2.00E-46	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
SMXD51_ORF_1013	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
SMXD51_ORF_1055	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
SMXD51_ORF_1276	40	583	6.00E-146	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
SMXD51_ORF_1280	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
SMXD51_ORF_1513	41	233	3.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
SMXD51_ORF_1702	97	242	5.00E-173	258	ErmC is a methyltransferase that catalyzes the methylation of A2058 of the 23S ribosomal RNA in two steps
SMXD51_ORF_1720	89	639	0	644	TetM is a ribosomal protection protein that confers tetracycline resistance
SMXD51_ORF_1722	98	436	0	442	TetL is a tetracycline efflux protein found in many species of Gram-negative and Gram-positive bacteria
UCC118_ORF_41	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
UCC118_ORF_61	50	498	3.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
UCC118_ORF_240	48	228	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
UCC118_ORF_676	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
UCC118_ORF_717	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
UCC118_ORF_1015	40	583	1.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
UCC118_ORF_1019	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
UCC118_ORF_1271	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
UCC118_ORF_1489	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
UCC118_ORF_1501	45	305	1.00E-79	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
UCC118_ORF_1583	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
UCC118_ORF_1825	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
UCC119_ORF_52	45	225	1.00E-62	236	vanRM is a vanR variant found in the vanM gene cluster
UCC119_ORF_216	48	228	2.00E-70	229	vanRF is a vanR variant found in the vanF gene cluster
UCC119_ORF_408	50	498	3.00E-171	493	LsaA is an ABC efflux pump expressed in Enterococcus faecalis
UCC119_ORF_594	46	418	2.00E-121	441	Mycobacterium tuberculosis murA confers intrinsic resistance to fosfomycin
UCC119_ORF_634	69	396	0	395	Sequence variants of Streptomyces cinnamoneus elongation factor Tu that confer resistance to elfamycin antibiotics
UCC119_ORF_933	40	583	1.00E-145	578	AdeC is the outer membrane factor of the AdeABC multidrug efflux complex
UCC119_ORF_937	45	648	0	772	PBP1a is a penicillin-binding protein found in Streptococcus pneumoniae
UCC119_ORF_1189	41	233	4.00E-54	233	vanRE is a vanR variant found in the vanE gene cluster
UCC119_ORF_1436	40	222	7.00E-47	224	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
UCC119_ORF_1447	45	305	1.00E-79	302	bcrA is an ABC transporter found in Bacillus licheniformis that confers bacitracin resistance
UCC119_ORF_1528	44	222	1.00E-61	227	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones
UCC119_ORF_1774	43	219	5.00E-53	233	MacB is an ATP-binding cassette (ABC) transporter that exports macrolides with 14- or 15- membered lactones

Table S9

gene	functional annotation	% identity	query coverage	e-value	query length
01M14315_ORF_1114	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
01M14315_ORF_1197	UDP-glucose_pyrophosphorylase	72.7	293	4.00E-151	290
778_ORF_1065	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
778_ORF_1147	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
866_ORF_1162	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
866_ORF_1246	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
ACS116_ORF_1278	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
ACS116_ORF_1363	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
AH4231_ORF_1274	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
AH4231_ORF_1357	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
AH4331_ORF_1249	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
AH4331_ORF_1331	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
AH4331_ORF_2194	cell_wall_surface_anchor_family_protein,_Plasminogen-_and_Fibronectin-binding_protein_B	70.3	101	9.00E-43	105
AH43324_ORF_1245	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
AH43324_ORF_1326	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
AH43348_ORF_1213	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
AH43348_ORF_1315	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
ATCC11741_ORF_1065	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
ATCC11741_ORF_1147	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG2753OB_ORF_1016	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG2753OB_ORF_1100	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG38008_ORF_1126	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG38008_ORF_1209	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG44481_ORF_1134	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG44481_ORF_1217	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG45735_ORF_1169	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG45735_ORF_1251	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG47171_ORF_1127	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG47171_ORF_1233	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
CCuG47825_ORF_396	polysaccharide_biosynthesis_protein_CpsF	73.63	91	5.00E-46	93
CCuG47825_ORF_1465	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG47825_ORF_2004	UDP-glucose_pyrophosphorylase	72.35	293	7.00E-151	290
CCuG47826_ORF_1150	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CCuG47826_ORF_1339	UDP-glucose_pyrophosphorylase	72.35	293	7.00E-151	290
CECT5713_ORF_1313	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
CECT5713_ORF_1425	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
cp400_ORF_1280	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
cp400_ORF_1413	UDP-glucose_pyrophosphorylase	73.04	293	4.00E-152	290
DSM18933_ORF_62	UDP-glucose_pyrophosphorylase	73.04	293	7.00E-153	300
DSM18933_ORF_219	polysaccharide_biosynthesis_protein_CpsF	74.5	149	7.00E-86	149
DSM18933_ORF_1644	ATP-dependent_Clp_protease_proteolytic_subunit	73.58	193	1.00E-104	196
DSM20492_ORF_1084	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
DSM20492_ORF_1168	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
DSM20492_ORF_1456	polysaccharide_biosynthesis_protein_CpsF	75	132	5.00E-73	132
DSM20554_ORF_1092	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
DSM20554_ORF_1174	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
DSM20554_ORF_1467	polysaccharide_biosynthesis_protein_CpsF	75	132	5.00E-73	132
DSM20555_ORF_1055	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
DSM20555_ORF_1138	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
GJ24_ORF_1096	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
GJ24_ORF_1189	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
gul1_ORF_1025	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
gul1_ORF_1107	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
gul2_ORF_1067	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
gul2_ORF_1150	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1040_ORF_1109	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1040_ORF_1192	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1042_ORF_1056	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1042_ORF_1141	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1044_ORF_1016	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1044_ORF_1100	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1045_ORF_1131	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1045_ORF_1214	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1046_ORF_1085	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1046_ORF_1167	UDP-glucose_pyrophosphorylase	73.04	293	4.00E-152	290
JCM1046_ORF_1500	polysaccharide_biosynthesis_protein_CpsF	75.17	149	9.00E-84	152
JCM1047_ORF_1287	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1047_ORF_1370	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
JCM1230_ORF_1032	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
JCM1230_ORF_1160	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
L21_ORF_1211	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
L21_ORF_1294	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
LMG14476_ORF_1187	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
LMG14476_ORF_1272	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
LMG14477_ORF_1171	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
LMG14477_ORF_1256	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NCIMB702343_ORF_1073	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
NCIMB702343_ORF_1158	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NCIMB702343_ORF_1488	polysaccharide_biosynthesis_protein_CpsF	77.14	140	7.00E-79	141
NCIMB8816_ORF_1018	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
NCIMB8816_ORF_1141	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NCIMB8816_ORF_1869	cell_wall_surface_anchor_family_protein,_Plasminogen-_and_Fibronectin-binding_protein_B	70	110	4.00E-47	110
NCIMB8817_ORF_1119	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
NCIMB8817_ORF_1201	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NCIMB8818_ORF_1316	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
NCIMB8818_ORF_1399	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NIAS840_ORF_276	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
NIAS840_ORF_577	polysaccharide_biosynthesis_protein_CpsF	75	132	5.00E-73	132

NIAS840_ORF_1603	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
Ren_ORF_1171	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
Ren_ORF_1266	UDP-glucose_pyrophosphorylase	72.35	293	4.00E-150	290
SMXD51_ORF_1498	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
SMXD51_ORF_1595	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290
UCC118_ORF_1257	ATP-dependent_Clp_protease_proteolytic_subunit	72.59	197	1.00E-105	197
UCC118_ORF_1367	UDP-glucose_pyrophosphorylase	72.7	293	2.00E-151	290