CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Development of methods for the microbiome analysis of formalin fixed paraffin embedded tissue specimens
Authors	Flores, Yensi
Publication date	2019-04-30
Original Citation	Flores Bueso, Y. A. 2019. Development of methods for the microbiome analysis of formalin fixed paraffin embedded tissue specimens. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2019, Yensi Alejandra Flores Bueso. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-15 14:04:04
Item downloaded from	https://hdl.handle.net/10468/10095



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

THE NATIONAL UNIVERSITY OF IRELAND, CORK

Coláiste na hOllscoile, Corcaigh

UNIVERSITY COLLEGE CORK

CancerResearch@UCC



Development of methods for the microbiome analysis of formalin fixed paraffin embedded tissue specimens

Thesis presented by:

Yensi Flores Bueso BSc, MSc, MBA

Under the supervision of

Mark Tangney BSc, PhD, MBA

for the degree of:

Doctor of Philosophy

April 30, 2019

Table of Contents

DECLARATION....5

ACKNOWLEDGMENTS....6

ABSTRACT....8

LIST OF ABBREVIATIONS....11

CHAPTER 1: 16

***INTRODUCTION*....16**

1. SYNTHETIC BIOLOGY IN THE DRIVING SEAT OF THE BIOECONOMY....17

Synthetic Biology R&D: Revolutionising Biotechnology 17

Is There Really A Biotech Revolution? The Markets Say Yes....21

Shaping the Bioeconomy: Synthetic biology influence on the biotech market and the bioeconomy....22

Concluding Remarks 25

Glossary 25

Online Resources 25

2. IN SITU BIOMOLECULE PRODUCTION BY BACTERIA; A SYNTHETIC BIOLOGY APPROACH TO MEDICINE 27

Bacterial-Produced Anti-Disease Agents....27

Bacteria as Region-Specific Colonisers....28

Synthetic Biology as a Technology....31

Cancer as an example indication....37

Synthetic biology approaches to improvement of bacterial agents and treatment strategies ..43

Treatment strategies 48

Regulatory agency aspects....51

Concluding Remarks 52

Online Resources 52

3. FFPE-INDUCED DNA DAMAGE; RELEVANCE TO MICROBIOME ANALYSIS 53

Introduction....54

The FFPE process 55

Formaldehyde interaction with nucleotides....56

Formaldehyde interaction with DNA....57

Formaldehyde-driven DNA-Protein Crosslinks (DPC) 58

The cellular milieu....59

Effect of tissue processing and storage on FFPE samples *61*
DNA Damage found in FFPE specimens *61*
FFPE Effects on Bacterial DNA *69*
DNA Repair *71*
CONCLUDING REMARKS 74
REFERENCES 75
CHAPTER 2: 98
***PROTOBLOCK - A BIOLOGICAL STANDARD FOR FORMALIN FIXED SAMPLES* 98**
ABSTRACT 99
INTRODUCTION 100
METHODS 102
1. *Preparation of Protoblocks* *102*
2. *Confirmation of cell content by microscopy* *103*
3. *DNA Analysis* *104*
4. *Murine models* *106*
5. *Bioinformatics and Statistical Analysis* *107*
RESULTS 109
1. *Protoblock generation and validation* *109*
2. *Protoblock for assessing bias introduced by sample prep methods* *111*
3. *Assessment of bacterial DNA integrity following FFPE* *113*
4. *Characterising contaminants in the FFPE and sequencing workflow* *116*
DISCUSSION 118
CONCLUSION 119
SUPPLEMENTARY MATERIAL 120
Troubleshooting/Technical Considerations *123*
REFERENCES 126
CHAPTER 3: 131
***CHARACTERISATION OF FFPE-INDUCED BACTERIAL DNA DAMAGE AND DEVELOPMENT OF A DNA REPAIR METHOD FOR METAGENOMICS & METATAXONOMICS* 131**
ABSTRACT 132
INTRODUCTION 133
METHODS 135
1. *Preparation of FFPE blocks* *135*
2. *DNA Analysis* *137*
3. *Optimising cross-link reversal* *139*

Verifying cross-link reversal strategy. *140*
4. DNA repair *140*
5. Bioinformatics and statistical analysis *142*
RESULTS 144
1. Characterisation of bacterial FFPE DNA damage *144*
2. Development of a DNA repair strategy *147*
Optimisation of decrosslinking *148*
DNA glycosylases reduce sequence alterations in FFPE DNA *150*
Development of an in vitro Base Excision Repair system *152*
Analysis of combined decrosslinking and BER treatment *154*
DISCUSSION 157
CONCLUSION 158
SUPPLEMENTARY MATERIAL 159
REFERENCES 167
CHAPTER 4: 174
DEVELOPMENT OF A NOVEL PROTOCOL FOR BACTERIAL DNA EXTRACTION FROM FFPE SAMPLES 174
ABSTRACT 175
INTRODUCTION 176
METHODS 181
1. Models *181*
2. DNA analysis *184*
3. Host depletion strategy *187*
4. Bacterial lysis strategy *190*
5. Integration of the protocol *190*
6. Validation of the protocol *194*
7. Bioinformatics *202*
RESULTS 204
1. Strategy for host DNA depletion *204*
2. Bacterial lysis strategy *208*
3. Integration of a Protocol for microbiome analysis of FFPE tissues *211*
4. Validation of the protocol by 16S sequencing *212*
DISCUSSION 221
CONCLUSION 226
SUPPLEMENTARY MATERIAL 227
REFERENCES 232

CHAPTER 5: 238

DISCUSSION 238

REFERENCES 243

DECLARATION

I hereby declare that I am the sole author of this thesis.

I authorise University College Cork to lend and photocopy this thesis to other institutions or individuals for the purpose of scholarly research.

Yensi Flores Bueso

ACKNOWLEDGMENTS

I am very grateful to the support of the Irish Research Council for funding my PhD. I also want to express my gratitude to my Supervisor, Dr Mark Tangney, for his amazing support, guidance, and encouragement throughout my PhD and MSc degrees. I will always be thankful and honoured to have been your student. I also want to appreciate my MSc supervisor Dr Kellie Dean, without whom, this would have only been in my dreams.

I want to thank all the members of the CancerResearch@UCC lab, who made the days happier and work enjoyable. Among all, I want to thank Dr Garret Casey, who has always keen to help with even the smallest things and Ms. Juliet Barry, who kindly and patiently trained me and helped me with all the histology workflow. I also want to appreciate the collaboration and kindness I received from members of my team, with whom I shared wonderful experiences. In particular, Ciaran Devoy, Stephen Buckley and Sidney Walker, who dared to travel across the world to a not very safe Honduras, to share their knowledge and inspire kids to pursue a research career. Among them, I have to say special thanks to Sidney, for all the support and positivism that kept me going despite so many failures. Thanks for answering email and messages at very odd hours and days. Also, apologies for interfering in many of your holidays.

I want to thank my good friends Jennifer Quinn and Kasia Komolibus. Thanks for keeping me alive, fed and sane this last year. Thanks for filling the hardest times, with moments of sympathy, laughs and wine! I also want to thank my friend Pramod for his inspiration, kindness and support.

I thank the support of my family, who despite the distance, are always there for me. Special thanks to my parents for their support, acceptance and incredible patience for putting up with me and all my crazy ideas. I thank the amazing brothers I have, without whom life would not have the safe meaning. Finally, I thank the person who has given me the inspiration, drive and strength to follow my dreams no matter the circumstance. Thanks Nana, for showing me what it is to live a meaningful life.

For my grandmother: Lucila Quan

ABSTRACT

For Synthetic Biology to reach its potential, it necessitates foundational knowledge of the organisms that can be engineered. The remarkable influence our microbiome has on our health status has made it a focus of attention for engineering possibilities aiming at its modulation. As the field of the human microbiome expands, it necessitates access to high-quality nucleic acid samples which are truly representative of the community of bacteria under study. Formalin-fixed, paraffin-embedded (FFPE) samples represent the most comprehensive collections of patient materials in hospital pathology archives. However, for this sample to become reliably accessible for microbiome studies, the effects of FFPE processing on bacteria must be considered.

Any sample processing method should be based upon specific study aims, target organisms and sample types. It is only through a holistic understanding of FFPE-induced changes to the bacterial cellular structure and its DNA content, that a reliable method can be developed It is hypothesised here that with a sample-prep workflow considering the effects of FFPE on bacterial cells, their DNA content and the overall contamination introduced, a reliable and reproducible analysis of the microbiome of FFPE samples could be achieved. As such, the overall aim of this thesis was to characterise FFPE induced changes to the bacterial cell walls/membranes and their DNA content, and with this information, to propose strategies for purifying and repairing DNA suitable for microbiome analysis, while also characterising the common contaminants found in samples processed in this manner.

To achieve this, an appropriate FFPE bacterial study model was first developed. With this in place, a thorough characterisation of the state of bacterial FFPE DNA was performed and strategies to reduce this damage assessed. Finally, to develop an appropriate method for bacterial DNA extraction from FFPE samples (unavailable at the time of writing), the state of the bacterial cell wall/membrane was assessed and strategies for a uniform bacterial lysis and host depletion evaluated.

Chapter 2 Describes methods for creating a mock bacterial FFPE block (Protoblock) that serves as a standard for FFPE samples. The Protoblock is a cell matrix which can be populated with cell types and numbers as desired, so as to resemble those of the FFPE tissue specimens. Its accuracy for representing bacterial load and cell architecture was validated by microscopy. With this model, the performance of the human gold-standard FFPE kit for microbiome analysis of FFPE samples was evaluated and found unsuitable for microbiome research. Additionally, the Protoblock permitted the characterisation of bacterial FFPE DNA, where it was found to be highly fragmented ($\bar{x}$ length = 143 bp), a poor PCR template (with a log-fold loss of amplifiable 200 bp fragments) and featured significant sequence alterations. Finally, this model also permitted the characterisation of contaminants originating from the FFPE process, the most common being Xanthomonadaceae, Pseudomonadaceae and Clostridiaceae.

Chapter 3 Makes a thorough investigation of the state of bacterial FFPE DNA in terms of PCR readability, formalin crosslinking, and the presence of sequence artefacts. Here, bacterial FFPE DNA was found to be highly fragmented, with a significant inverse correlation between fragment size and PCR recovery and a log-fold reduction between the recovery of 200 bp and 500 bp fragments. It was also evident that 95-97% of DNA present in these samples was crosslinked and that the most evident sequence artefacts were those derived from oxidative damage. Two strategies to reduce this damage were investigated. (1) An optimised decrosslinking procedure (10 ^{o}C lower than current methods) significantly reduced sequence artefacts generated by high-heat incubation. (2) The in vitro reconstitution of the Base Excision Repair pathway targeting oxidative DNA damage, using FPG and Endo VIII DNA glycosylases. Samples treated with both strategies showed a 3X increase in fragment length and a significant reduction in sequence chimeras and SNPs, leading to a significant improvement in sequencing readability.

Chapter 4 Investigates the state of the bacterial cell wall/envelope and mammalian membrane to assess the state of their permeabilisation in FFPE samples._In this chapter, mammalian and Gram-negative bacterial cells were found to be impermeable to molecules with dimensions of 3-5 nm. A host depletion strategy was devised using a combination of Saponin and DNAse (Benzonase). It was also found that FFPE bacterial cells require a lysis strategy, and the use of a mix of bacterial-lytic enzymes was found to provide a uniform cross-taxa bacterial lysis. The integration of different treatments was achieved using 0.2 µm CA filtering columns between treatments. The collection of methods developed were tested by 16S rRNA gene sequence analysis of protoblocks, murine FFPE faeces and human breast tumour samples. The collection of methods provided an overall increase in recovery of 16S PCR amplicons, a higher uniformity in bacterial lysis, and a higher bacterial to host DNA ratio in high biomass models. However, these improvements were obscured for low biomass samples, where contaminants dominated the sequencing reads.

It is concluded from this work that to unlock the potential of FFPE specimens for the microbiome field, a full dedicated workflow, comprising not only sample-prep, but also QC, 16S PCR and 16S sequencing, needs to be in place. This workflow should be directed by a robust QC system. In addition, a database for known FFPE derived common contaminants is essential to inform future strategies for the biological removal of contaminants from these samples.

LIST OF ABBREVIATIONS

AHL	N-3-oxohexanoyl-L-homoserine lactone
APC	Antigen presenting cells
ASV	Amplicon Sequence Variants
AP	Apurinic/ Apyrimidinic
BER	Base Excision Repair pathway
BLS	Bacterial lysis solution
bp	Base pair
BSA	Bovine Serum Albumin
CA	Cellulose Acetate
CD	Cytosine deaminase
CDx	Cluster of differentiation
CFU	Colony forming unit
CRISPR	Clustered regularly interspaced short palindromic repeats
CTLA-4	cytotoxic T-lymphocyte antigen-4
Cys	Cysteine
dA	deoxyadenine
DC	Dendritic cells
DCL	Decrosslinking solution
dC	deoxy Cytosine
dG	deoxy Guanine
dI	deoxy Inosine
dT	deoxy thymine
dU	deoxy Uracil
dH	Dihydro
diOH	dihydroxyl
dNTPS	Deoxy nucleotides
DPC	DNA Protein crosslinks

dRP	Deoxyribose phosphate
ds	Double strand
DT-N	in house method + host DNA depletion **without** tissue dissociation
DT-P	in house method + host DNA depletion **with** tissue dissociation
DTT	Dithiothreitol
EB	Elution buffer (10 mM Tris-HCL)
EDTA	Ethylenediaminetetraacetic acid
Endo IV	Endonuclease IV
Endo VIII	Endonuclease VIII
FAB	Fastidious Anaerobe Medium
fapy-dG	fapy-deoxyguanine
FF	Formalin fixed
FMT	Faecal microbiota transplantation
FFPE	Formalin-fixed, paraffin-embedded
FPG	Formamido-pyrimidine DNA glycosylase
G-	Gram-negative
G+	Gram-positive
GF	Germ-free
GIT	Gastrointestinal tract
GuHCL	Guanidine Hydrochloride
h	Height
h	Hour
H&E	Haematoxylin & Eosin
hAAG	Human Alkyl Adenine DNA Glycosylase
HCl	Hydrochloride
HCOH	Formaldehyde
HD	Host Depletion
HDB	Host Depletion Buffer
HDS	Host Depletion Solution

His	Histidine
hm-	Hydroxymethyl
HRM	High resolution melt analysis
IHN	in house **without** host depletion
IHP	in house **with** host depletion
IPO	Initial Public Offering
IPTG	Isopropyl β- d-1-thiogalactopyranoside
Kbp	Kilobases
KDa	Kilodalton
LB	Luria-Bertani medium
LPS	Lipopolysaccharide
Lys	Lysine
M	Molar
MBp	Mega base pair
me	Methyl
MgCl2	Magnesium Chloride
min	Minutes
ml	Millilitre
mM	Milimolar
mm3	Cubic millimetre
MHC	Mayor Histocompatibility Complex
MRS	De Man, Rogosa and Sharpe Medium
MW	Molecular weight
N	Nitrogen
NaCl	Sodium Chloride
NAD+	Nicotinamide adenine dinucleotide
NF	Non-fixed
ng	Nanogram
NGS	Next-Generation Sequencing

NH	Amino group
nm	Nanometre
oC	Degrees Celsius
OD600	Optical density at 600 nanometres
OH	Hydroxy
OM	Outer-membrane
OTU	Operational taxonomic units
P	Phosphate
PBS	Phosphate Buffer Saline
PCR	Polymerase Chain Reaction
PD	Phosphodiester bond
PDL1	protein 1/programmed cell death 1 ligand 1
pH	Hydrogen potential
PNK	Polynucleotide Kinase
Pol I	Polymerase I
PVDF	Hydrophilic Polyvinylidene Fluoride
Q Protocol	QIAGEN QIAMP DNA FFPE Tissue Protocol
QF	Quality Filter
qPCR	Quantitative PCR
R&D	Research & Development
RCM	Reinforced Clostridial medium
RT	Room Temperature
sec	Seconds
RT	Short-chain fatty acids
SNP	Single Nucleotide Polymorphism
ss	Single strand
T4-PNK	T4 Polynucleotide Kinase
TAA	Tumour-associated Antigens
TAM	Tumour-associated Macrophages

TB	Test Buffer
TBS	Tris Buffer Saline
TDG	Thymidine DNA Glycosylase
TDS	Tissue dissociation solution
Th	T helper cells
Tm	melting temperature
ΔTm	melting temperature difference
Treg	Regulatory T-cell
Trp	Tryptophan
U	Units
UDG	Uracil DNA Glycosylase
ug	Micrograms
ul	Microliter
um	Micrometre
uM	Micromolar
V	Volume
VC	Venture capital
WGS	Whole Genome Sequencing
X	Times
x g	Times gravity (relative centrifugal force)
8-oxo-dA	8-oxo-deoxyadenine
8-oxo-dG	8-oxo-deoxyguanine
α-	anti-
Δ	Delta = Change
□	Average
S	Sum
~	Approximate

CHAPTER 1:

Introduction

Sections of this chapter have been published as:

Flores Bueso, Y. and Tangney, M. (2017). **Synthetic Biology in the Driving Seat of the Bioeconomy.** *Trends in Biotechnology*, 35(5), pp.373-378. Impact Factor – 13.75

Flores Bueso, Y., Lehouritis, P. and Tangney, M. (2018). **In situ biomolecule production by bacteria; a synthetic biology approach to medicine.** *Journal of Controlled Release*, 275, pp.217-228. Impact Factor – 7.91

O'Connor, H., MacSharry, J., Bueso, Y. F., Lindsay, S., Kavanagh, E. L., Tangney, M., Clyne, M., Saldova, R., McCann, A. (2018). **Resident bacteria in breast cancer tissue: pathogenic agents or harmless commensals?** *Discov Med*. **26**(142): p. 93-102. Impact Factor – 2.38

1. Synthetic Biology in the driving seat of the Bioeconomy

Synthetic biology is revolutionising the biotech industry and is increasingly applied in previously unthought-of markets. We discuss the importance of this industry to the bioeconomy and two of its key factors: the synthetic biology approach to R&D, and the unique nature of the field's carefully designed, stakeholder-inclusive, community-directed evolution.

Synthetic Biology R&D: Revolutionising Biotechnology

Synthetic biology is a young field that emerged from the convergence of biosciences, information technology and engineering. Since its coinage, synthetic biology has evolved as an umbrella term, defined by a conceptual framework, aiming at the rational design of biological systems to attain useful products. This is sought through the integration of engineering principles at the core of the R&D cycle and replacing ad hoc and serendipitous practices characteristic of traditional biotechnology. This results in a more reliable and robust industry, compatible with automation and scalability, while also upgrading its capabilities to carbon-neutral, simplified production systems, with a wider scope of products [1].

Synthetic biology has become a global enterprise, expanding to over 40 countries, with almost 700 organizations conducting synthetic biology research, funded by over 530 funding agencies [2]. The promising breakthroughs and disruptive technological advances delivered by synthetic biology are evidence of the support placed by governments and funding bodies, who also fostered programmes that facilitate routes to market and the creation of a thriving heterogeneous community [3]. As a consequence, the scope of the biotech industry and its market has been revolutionised to a scale that required a new dimensional definition - '*The Bioeconomy'* [4].

Here, we adopt the OECD Bioeconomy definition - the share of the economy delivered by biotechnology - although other definitions also include activities transforming bio-resources. Nowadays the Bioeconomy is included in economic roadmaps of many nations and regions, where synthetic biology is a key technology, enabler of the global transition to a bio-based economy. This transition is proposed as the ultimate solution

to sustainably fulfil the current and future food, health and energy demands of a growing global population, where already scarce natural resources are challenged by constrains of climate change [4].

Turning synthetic biology into a Global Enterprise (Figure 1)

Promising projects emerging in the mid-2000's encouraged the launch of major EU and US federal initiatives. Among them, in 2006, the NSF funded $40 million to initiate the SynBERC (Synthetic Biology Engineering Research Centre), a multi-institutional research centre fast-tracking the commercialisation of synthetic biology products, and since then, the NSF has allocated almost $140 million to synthetic biology research [5]. The same year, after a presidential mandate to develop a biofuel economy, the US Department of Energy (DOE) announced a $1bn investment in a multi-institutional consortium to progress each stage of biofuel production (from basic research and enabling technologies, to crops and microbes) with more than $400 million allocated to synthetic biology related research [i].

The US government has invested approximately $500 million–$1bn in synthetic biology research since 2005, with a marked 200% increase in 2010, the year when the US Defence Advanced Research Projects Agency (DARPA) launched the precursors of its Living Foundries - ATCG and 1000 molecules programs [ii] - levering bioengineering capabilities to manufacturing platforms. The NIH has funded more than $50 million during 2005-2010 and $20 million 2014-2019 the Genomes-to-Natural Products program [3].

Meanwhile in the EU, a gross figure of €450 million in synthetic biology funding was reported from 2004-2013 [iii]. The UK began funding synthetic biology activities in 2007, and since then has become the world's second-most active nation in synthetic biology activities and the European leader, investing over £300 million in synthetic biology activities. Unlike the US, UK synthetic biology programmes have been developed under a unified strategy, as detailed in its roadmap, focusing on developing a robust research community with strong links to industry. In 2016, the UK developed a new Strategic Plan: 'BioDesign for the BioEconomy', aiming at higher impact for their synthetic biology market [1]. Overall, centres performing synthetic biology

activities have been reported in 17 countries in Europe. On the other side of the world, China published its synthetic biology roadmap in 2010, allocating 260 million Yuan ($36 million) per annum [5]. In 2016, the Human Genome Project write-up was launched with an initial budget of $100 million, sourced from private and public organisations worldwide. However, its total cost is expected to exceed $3bn [6].

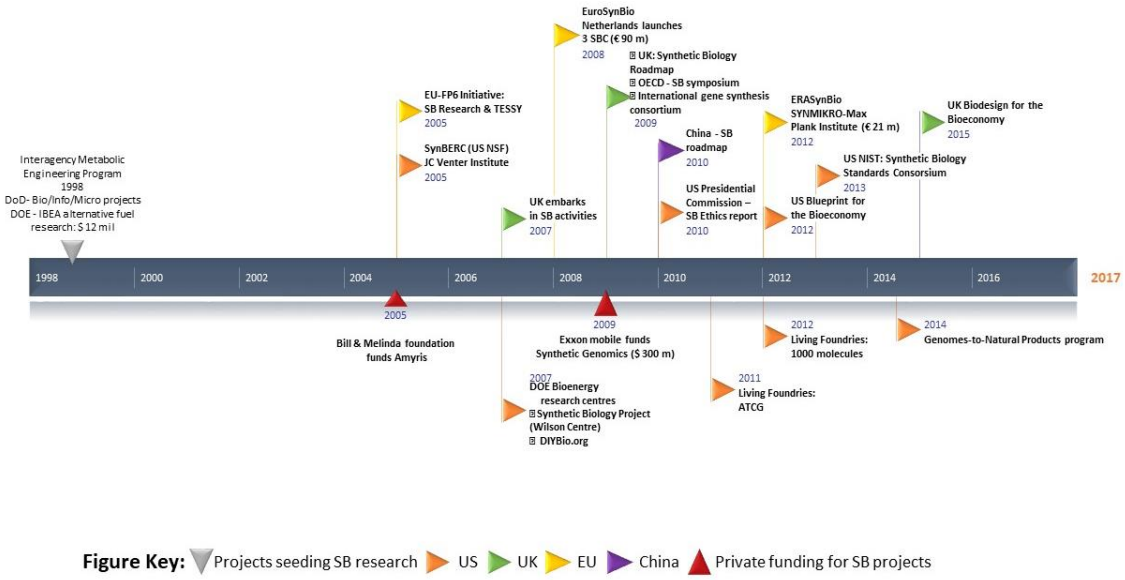


Figure 1. Initiatives driving synthetic biology expansion and its adoption as an industrial technological platform

Synthetic biology fosters disruptive technological advances (Figure 2)

DNA Sequencing – As an outcome of the thousand-dollar genome project, next-generation sequencing (NGS) and third-generation sequencing platforms were conceived. These revolutionary technologies have increased productivity more than 500-fold, changing the sequencing economics and seeding an industry that since 2007 has outpaced Moore's Law, yielding a 10,000-fold price decrease for a human genome relative to the cost in 2004 [iv].

DNA Synthesis and Assembly – Over the past decade, traditional *de novo* DNA synthesis methods have been significantly improved, but the introduction of novel microchip-based DNA synthesis strategies has represented a truly disruptive

technology in this industry, as it enables miniaturised, in-parallel and automated production, increasing throughput and efficiency, and increasing productivity more than 700-fold. These advances have facilitated the synthesis of gene-size DNA fragments and prompted a 10^4-10^6 decrease in price for oligonucleotide, and a 100-fold decrease for gene, synthesis. Assembly of longer DNA constructs (>2 Kb) is now possible through novel high fidelity *in vitro* enzymatic assembly methods that are also inexpensive and suitable for automated systems[7].

Genomic Engineering – The advent of the CRISPR/Cas9 as a genome editing technology in 2013 marked the beginning of a new genome-engineering era. Since its publication in 2013, it has proven effective in a multitude of organisms, including humans. Its superior efficacy and precision, coupled with its simplicity and low cost (< $100), has revolutionised the genomic engineering arena, enabling its widespread adoption in research and industry, by both trained scientists and amateurs [v].

Mathematical modelling – elementary tool for the rational design of robust and complex synthetic biology systems. It enables abstraction and increases the speed and reliability of building synthetic biological devices and systems, by reducing the amount of time-consuming, expensive and unpredictable wet-lab experiments. By increasing the reliance of a project in *in silico* models, it accelerates innovation and reduces costs [8, 9].

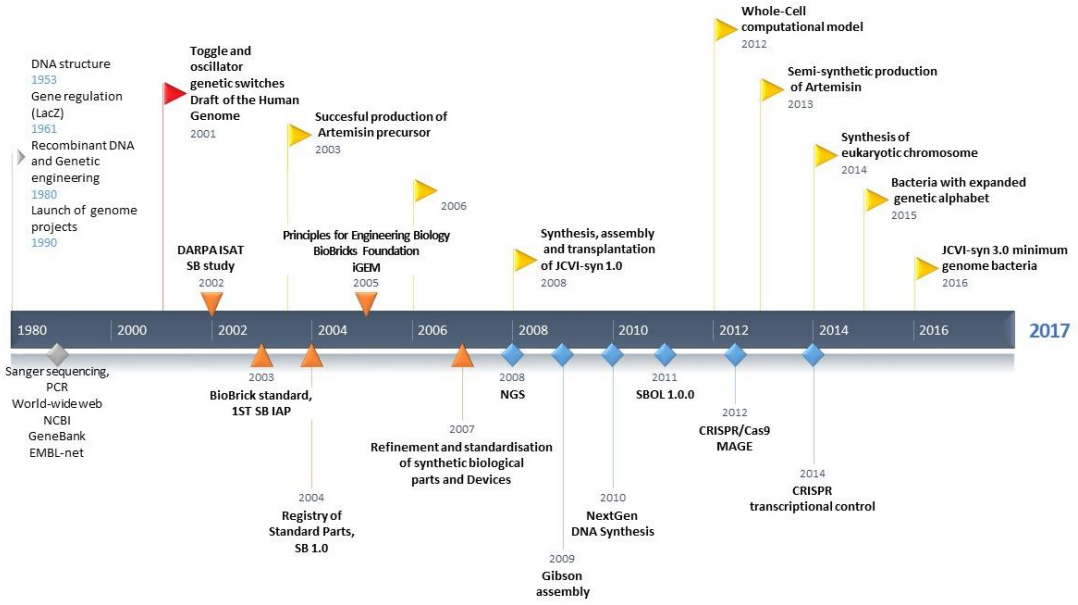


Figure 2. Scientific breakthroughs, advances in enabling technologies, and milestones that have built synthetic biology as a discipline

Is There Really A Biotech Revolution? The Markets Say Yes

Current reports estimate that biotech contributed $324bn to the US bioeconomy by 2012 (more than 2% of US GDP), with an annual 10% growth over the last decade [10]. The EU has estimated a value of €2 trillion – although including activities beyond biotech [vi]. Similarly, the UK has estimated its bioeconomy to be £150bn, predicting a growth of £40bn over the next decade [vii].

Biotech's expansion has also been perceived in the public and private markets, with the longest and largest expansion in biotech history beginning in 2009 and peaking in 2015. Since 2013, more than 224 companies launched to the public market with Initial Public Offerings (IPOs) that created a market value of $95bn. This unprecedented period has defined the longest, more prolific IPO window (period with more than four IPO per month) in biotech history [11]. In 2015, Biotech companies raised an unprecedented $71bn, with 58% ($41.3bn) corresponding to innovation capital (raised by companies with <$500 million in revenues), featuring a record-high $11.8bn in venture capital funds (30%), and also record-high $3.5bn early-stage funding (235 series A and seed funding) [12]. Moreover, despite the market's contraction entered in 2016 (typical of the biotech market cyclic behaviour), the IPO window remained open, featuring a high proportion of early-stage companies, who raised 30% in valuations prior to investment [viii]. Similarly, venture funding to biotech remained strong, with US VC firms investing ~$7bn (well above the $4.2bn historic average), while average start-up investment doubled from $7.5 to $15 million per deal [ix].

Overall, despite the markets' volatility, the biotech industry has kept a steady expansion, and the NASDAQ biotech index is still performing 160% of five years ago. The industry is providing a rich pipeline of products, with high capitalisation of its players (high valuations), higher flow of funds in the equity market (more IPO's than any other industry sector) and a larger participation of earlier-stage players, all of which is characteristic of a prolific industry [11].

Shaping the Bioeconomy: Synthetic biology influence on the biotech market and the bioeconomy

Despite its youth, synthetic biology already has a significant market participation, valued at $2.7bn in 2013, $3.9bn in 2016, growing at a Compound Annual Growth Rate (CAGR) of 24.4%, and expected to reach ~$11.4bn by 2021 [13]. However significant, this valuation is conservative, since its contributions are not confined to any readily-measured biotech industry segment, but benefit the overall industry. Synthetic biology has expanded the biotech industry by enabling its integration in other industries (e.g. Tech industry - DNA data archives [x], Nano-motors and molecular machines [xi]); attracting new players through easy-to-use and inexpensive tools (Amino Labs [xii], Bento Labs [xiii]) and enabling novel and sophisticated production systems and products. Overall, synthetic biology is an innovation platform driving the bioeconomy expansion, whose contributions go beyond research, and include the development of social and community-based initiatives facilitating its acceptance and integration by industry, society, governments and markets [1, 4, 5].

1. Accelerating the R&D cycle: The initiatives supporting SynBio prioritised the advancement of enabling technologies. Some of which have advanced the scope of numerous research areas in industry and academia, rendering a potential economic impact of between $700bn and $1.6 trillion per year by 2025 [14]. Beyond these technological advancements, synthetic biology transformed conventional R&D cycles (figure 3) in the biotech industry by integrating to its core, the following approaches that improve reliability, speed and costs: (1) *in silico* modelling (through abstraction) that reduce trial-and-error approaches. (2) Public repositories of standardised genetic components, enabling sharing of parts that can be reused and optimised for different purposes [8, 15]. (3) Decoupling R&D projects from manufacture (such as DNA synthesis), where research efforts and scaled manufacturing can be performed simultaneously by different entities [xiv]. (4) Automation and scaling [xv]. This framework advances innovation by providing a common language and infrastructure encouraging collaboration, enabling in-parallel work, and reducing the dependence on highly trained specialists and expensive and time-consuming lab work [xvi].

2. Fostering investment & entrepreneurship: By adopting this framework, biotech projects have become more reliable and feasible endeavours, which are more attractive for investment. Biotech now has an extensive investor base comprising recognised investor firms, among these, renowned Tech firms such as Google Ventures (GV) [xvii], and Y Combinator [xviii]. In addition, synthetic biology has more than 20 dedicated business incubator programmes (supported by industry, government, or VC firms) among them: LABS (Singularity University), IndieBio (US) and RebelBio (Ireland) [xix]. Undoubtedly, the landscape of biotech investment has changed, the overall marked increase in net capital flow and proportion of innovation and early-stage capital, highly supported by Corporate VC firms implies a structural change in the markets fostering entrepreneurship, as evidenced by funding raised by synthetic biology firms [xx] [11, 12]. Innovation is also promoted by increased accessibility and the numerous public-funded initiatives that recruit young talent to the field (from high-school to university undergraduates) through community-building activities such as LEAP [xxi], BioBuilder [xxii] and iGEM competition [xxiii], which has recently trained more than 25 thousand students, providing the workforce with key industry and entrepreneurship skills. These programs create an environment encouraging the generation and exchange of ideas that may develop into novel applications that expand the scope of synthetic biology and its markets [xxiv].

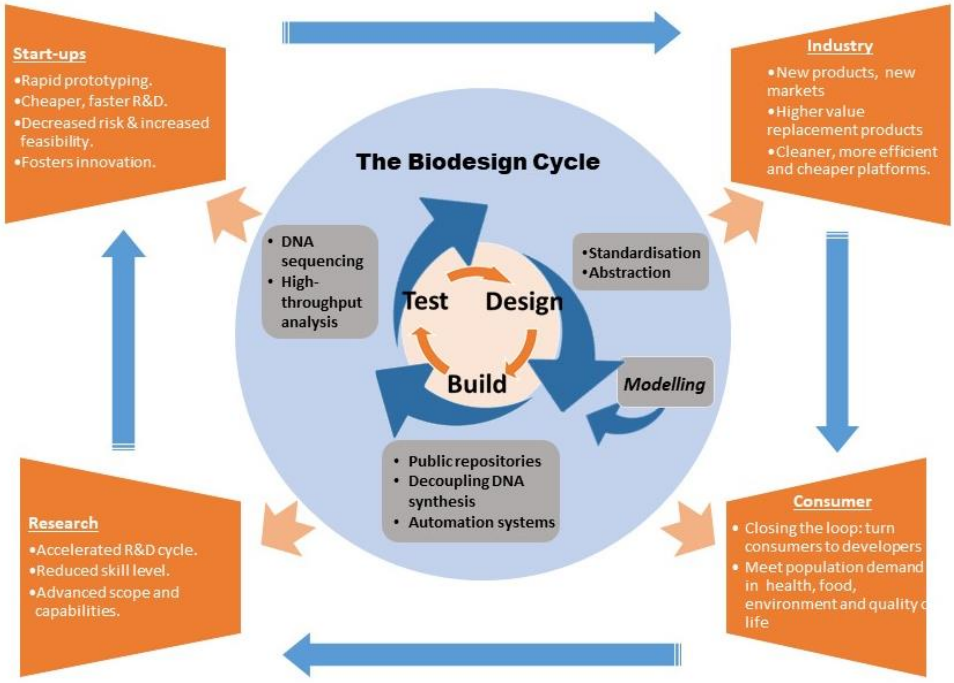


Figure 3. Synthetic Biology R&D cycle

3. Enabling new products: The novelty fostered by these initiatives and the increased capabilities of its technological contributions has expanded the scope of synthetic biology far beyond traditionally biotech-reliant industries (e.g. biopharmaceuticals). Novel products that were never previously considered are now emerging for less-saturated or less-restrictive markets, such as cosmetics, clothing, materials, nutrition, education and others, which, coupled with shorter R&D cycles, accelerates the pace for their launch to market [xxv]. Synthetic biology applications are now conquering new markets, shaping the bioeconomy by integrating it with industry sectors that are not accounted for in traditional biotech. Therefore, synthetic biology participation in the global market is not restricted to what is currently defined as biotech. A 2016 survey, reported more than 350 synthetic biology dedicated firms across US and EU (in 16 different industries) raising over $3.3bn between 2009-2015. That same year, 190 US SynBio companies raised $830 million [xxvi]. See a listing of companies at SynBioProject.org [xxvii].

4. Replacing traditional industrial processes: The technological advances brought by synthetic biology have enabled projects delivering high-value products in the traditional biotech sector. In the energy sector, synthetic biology has enabled the scalable production of biofuels and petroleum derivative products, which use carbon-neutral feedstock, improving its sustainability and ecological impact. Equally, the diagnostics industry has benefited from synthetic biology tools, applying them to multiple novel molecular and/or microfluidics diagnostics platforms that are now revolutionising the industry. Synthetic biology has also been adopted in the pharmaceutical industry, enabling drug discovery and the creation of more effective, safer and cheaper new generation drugs. Synthetic biology has been rapidly adopted for vaccine development, as it reduces the time for development significantly. It has also been adopted to replace older production methods, for more efficient and cheaper, one-step production systems that are more sustainable and harmonious with the environment. Similarly, other industries, in particular the chemical industry, have adopted synthetic biology to replace older production systems [10, 13].

Concluding Remarks

It remains to be seen if synthetic biology promises will be realized more than the earlier biotech hopes. The synthetic biology community, from the outset, placed much attention on bioeconomy aspects (product development needs, routes to market etc.) in order to avoid commercial failures observed with traditional biotech. Overall, the synthetic biology community-directed evolution approach makes it unique, and increasing success stories may lead to future recognition of this 'way to do business' as a game changer in scientific technology development.

Glossary

Diamond v Chakrabarty: court ruling enabling patents on GMOs

Initial Public Offering (IPO): The act of offering the stock of a company on a public stock exchange for the first time.

IPO Open Window: period with more than 4 IPO per month, indicative of market strength

Moore's law trend: A prediction made in integrated circuits, where the number of transistors in a chip will double every 2 years, while keeping the same price.

NASDAQ biotech index: is a stock market index for NASDAQ-listed companies, which is the second largest stock market.

Synthetic Biology (European Commission): The engineering of complex biological systems with novel functions, done in a rational and systematic matter, at all levels of hierarchical structures (molecules, cells, tissues, and organs)

Online Resources

i. https://energy.gov/articles/doe-provides-30-million-jump-start-bioenergy-research-centers
ii. http://www.darpa.mil/program/living-foundries
iii. http://www.evolva.com/wp-content/uploads/2016/01/EU-Synbio-Vision.pdf
iv. http://www.nature.com/news/technology-the-1-000-genome-1.14901
v. http://www.nature.com/news/crispr-the-disruptor-1.17673
vi. http://www.bioeconomyalliance.eu/node/83

vii. (https://connect.innovateuk.org/documents/2826135/31405930/BioDesign+for+the+Bioeconomy+2016+DIGITAL+updated+21_03_2016.pdf/d0409f15-bad3-4f55-be03-430bc7ab4e7e

viii. https://lifescivc.com/2016/07/biotechs-paradox-robustly-valued-highly-active-seemingly-terrible-ipo-market/

ix. https://lifescivc.com/2017/01/crystal-ball-gazing-biotech-predictions-2017/

x. http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496

xi. http://www.nature.com/news/the-tiniest-lego-a-tale-of-nanoscale-motors-rotors-switches-and-pumps-1.18262

xii. http://www.amino.bio/

xiii. https://www.bento.bio/

xiv. http://www.aiche.org/resources/publications/cep/2016/september/sbe-supplement-synthetic-biology-rewriting-dna-synthesis

xv. http://www.nature.com/news/the-automated-lab-1.16429

xvi. https://techcrunch.com/2015/09/28/synthetic-biology-is-not-just-good-its-good-for-you/

xvii. https://www.gv.com/portfolio/#life

xviii. https://www.ycombinator.com/biotech/)(http://www.nature.com/news/start-up-investor-bets-on-biotech-1.15096

xix. http://www.nature.com/news/young-scientists-ditch-postdocs-for-biotech-start-ups-1.20912

xx. http://www.nature.com/news/synthetic-biology-lures-silicon-valley-investors-1.18715

xxi. http://synbioleap.org/

xxii. http://biobuilder.org/

xxiii. http://igem.org/

xxiv. http://www.sciencemag.org/careers/2015/06/training-synthetic-biology-jobs-new-bioeconomy

xxv. http://www.nature.com/news/synthetic-biology-firms-shift-focus-1.14602

xxvi. http://www.synbioproject.org/cpi/companies/

2. In situ biomolecule production by bacteria; A synthetic biology approach to medicine

The ability to modify existing microbiota at different sites presents enormous potential for local or indirect management of various diseases. Because bacteria can be maintained for lengthy periods in various regions of the body, they represent a platform with enormous potential for targeted production of biomolecules, which offer tremendous promise for therapeutic and diagnostic approaches for various diseases. While biological medicines are currently limited in the clinic to patient administration of exogenously produced biomolecules from engineered cells, *in situ* production of biomolecules presents enormous scope in medicine and beyond.

The slow pace and high expense of traditional research approaches has particularly hampered the development of biological medicines. It may be argued that bacterial-based medicine has been 'waiting' for the advent of enabling technology. We propose that this technology is Synthetic Biology, and that the wait is over. Synthetic Biology facilitates a systematic approach to programming living entities and/or their products, using an approach to Research and Development (R&D) that facilitates rapid, cheap, accessible, yet sophisticated product development. Full engagement with the Synthetic Biology approach to R&D can unlock the potential for bacteria as medicines for cancer and other indications.

In this review, we describe how by employing Synthetic Biology, designer bugs can be used as drugs, drug-production factories or diagnostic devices, using oncology as an exemplar for the concept of *in situ* biomolecule production in medicine.

Bacterial-Produced Anti-Disease Agents

In '*ex vivo*' settings (industrial fermentation), engineered bacteria have long been used to produce recombinant proteins, such as insulin, human growth hormone and others [16, 17]. More recently, precedents have been set for bacterial production of small molecules and chemical entities for pharmaceutical uses [18, 19]. The commercial production of semi-synthetic artemisinin is frequently held up as the first demonstration of the potential of synthetic biology for the development and

production of pharmaceutical agents [18]. *E. coli* has been the bacterium of choice for the majority of agent production systems to date, although the range of bacterial genera is recently increasing with advances in engineering technology, and the capacity of different genera to provide more optimal agent production depending on the agent [20]. Given that *E. coli* and other bacteria can naturally, or be induced to, colonise different parts of the body, we ask if there is potential to 'skip the middle man', where the producing bacteria themselves may represent the final 'drug' product for administration to patients. In this context, the bacteria act as *in situ* drug producing 'biofactories', with the intervention focused at the site of pathology.

Bacteria as Region-Specific Colonisers

The microbiome research field has exploded in recent years, and while originally primarily focussed on bacterial colonisation of the gastrointestinal tract (GIT), research has expanded to various regions of the body, with characterisations of the microbiota of humans, animals, insects and non-living locations. 'Tract' regions of the human body, such as the vaginal and oral tract, feature distinct microbiota [21-23], and the microbiome of the skin, the largest organ of the body, is increasingly characterised [24]. The growing body of evidence supporting associations between the human microbiome and our health has drawn significant attention. The ability to modify existing microbiota at different sites presents enormous potential for local or indirect management of various diseases (Figure 1).

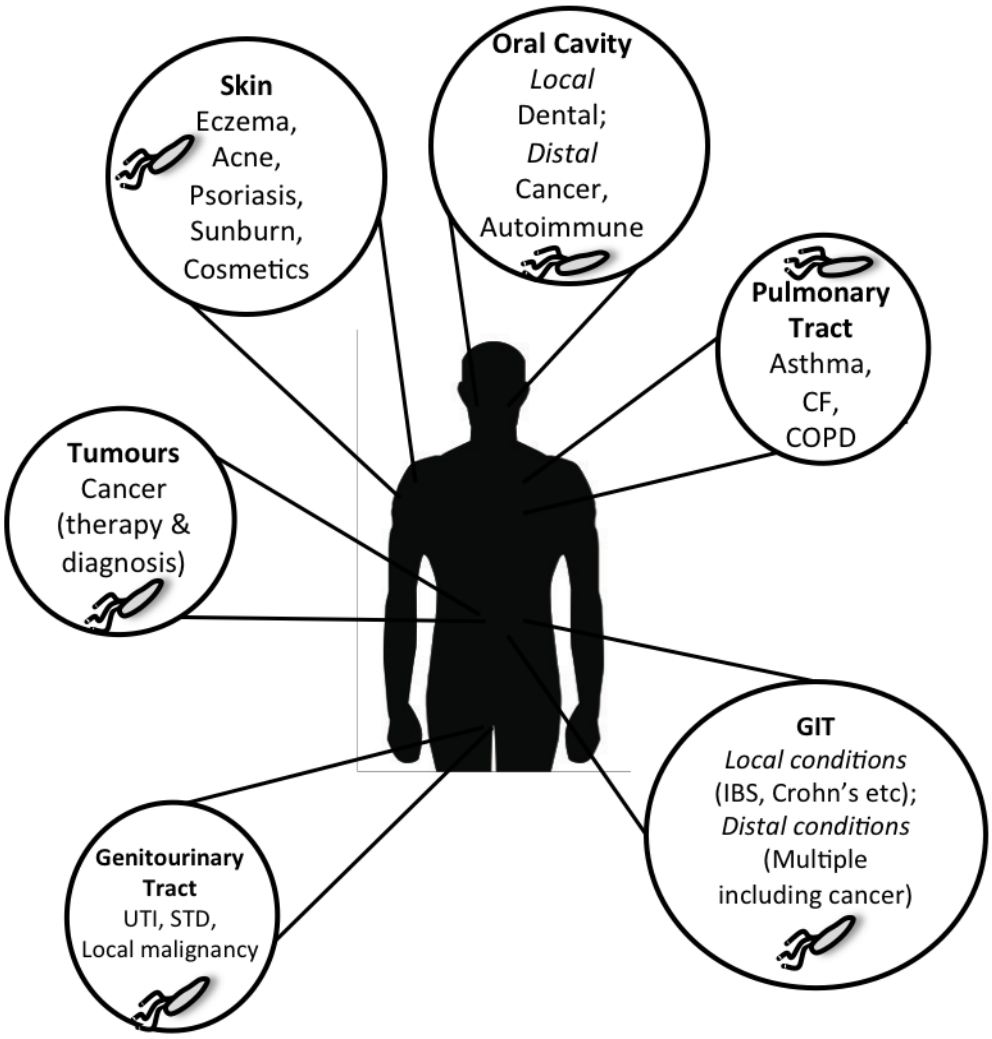


Figure 1. Example regions of the body where bacteria can be induced to colonise.
Sample conditions representing treatment targets for local bacteria are indicated for each location.

While the ability to induce growth of different bacteria in the GIT (via oral administration of probiotics) is widely known, there are precedents for artificial inoculation of other body sites. Table 1 shows a selection of examples of biomolecule production from bacteria at different body sites, examples of these are represented in Figure 2. In addition to supplementing the microbiome of more well described tract regions with engineered bacteria, targeting of solid tumours by this strategy is also under development.

Various studies have shown that tumours support the growth of different bacterial species, and many clinical and preclinical studies are underway to effect tumour-specific therapies through administration of engineered bacteria (see later). In addition

to these directly-acting therapies, associations between the nature of cancer patients' gut microbiota and tumour progression have been established [25]. For example, recent research in experimental cancer models has revealed that gut bacteria may influence the outcome of chemotherapy or immunotherapy indirectly via influencing the immune system [26, 27].

Table 1. Examples of in situ bacterial products in development for various diseases by body site

Company/ Product	Technology	Target Indication	Stage of develop-ment	Source
GIT				
ActoBiotics	*Lactococcus lactis in situ* production of cytokines, enzymes, hormones, and monoclonal antibodies	Allergic diseases, type 2 diabetes, autoimmune disorders (celiac disease; type 1 diabetes)	Clinical & preclinical	https://www.dna.com/Technologies/ActoBiotics
Synthetic Biologics (Ribaxamase)	*Lactococcus lactis in situ* production of therapeutic protein	*C. difficile infection and antibiotic-associated diarrhoea*	Clinical (Phase 2)	http://www.syntheticbiologics.com
Synlogic	Programming of the local microbiome metabolism: Probiotic bacteria with circuits that sense the patient's GIT environment regulate metabolic pathways	Inflammation, metabolism, oncology	Preclinical	http://www.synlogictx.com
Advaxis	*Listeria monocytogenes* delivery of Tumour-Associated Antigens to mucosal immune cells	Cancer (Cervical, Prostate, Breast)	Clinical (Phase 3; Phase 2)	http://www.advaxis.com
Oral cavity				
ActoBiotics AG013	*Lactococcus lactis in situ* production of TreFoil Factor-1	Oral mucositis	Clinical (Phase 1b)	[28]
Skin				
AOBiome	Ammonia-oxidizing bacteria (*Nitrosomonas*)	Acne, Eczema, Wound healing, Thermo regulation, Hypertension	Clinical (Phase 2; Phase 1). Preclinical	http://www.aobiome.com
TopgeniX	Platform technology for enduring application of natural compounds by skin microbiome	Sun protection, skin health, cosmetics	Preclinical	http://www.topgenix.com
Tumours				

Multiple (see this review)	Tumour-selective bacterial production of biomolecules	Any solid tumour	Clinical (Phase 2). Preclinical	[29]

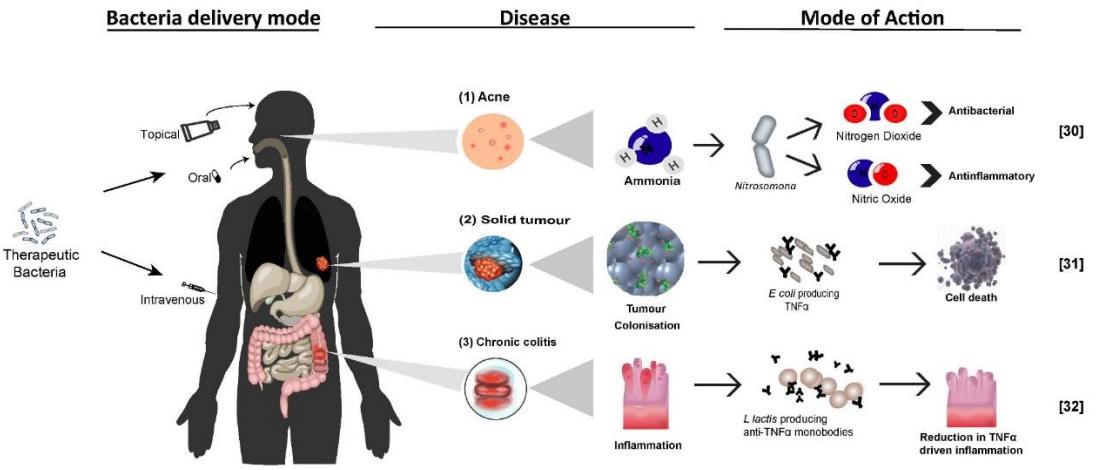


Figure 2. Illustration of in situ bacterial products, where:

(1) *Topical application of Nitrosomona eutropha oxidises ammonia into nitrogen dioxide (antibacterial) and nitric oxide (anti-inflammatory), preventing and treating acne [30].* ***(2)*** *Intravenously administered E coli MG1655 colonises solid tumours and delivers TNFα antibodies, impeding tumour growth [31].* ***(3)*** *Orally administered L. lactis delivers TNFα monobodies to the Colon significantly reducing inflammation in a chronic colitis model.[32]*

Synthetic Biology as a Technology

Synthetic Biology is an evolving discipline focused on engineering biological systems for global needs, representing an umbrella term that covers many approaches aimed at bestowing biological entities with novel functions or replicating biological functions outside a cell [33]. Synthetic biology aims at the rational design of biological systems by integrating engineering principles (standardisation, modularity, abstraction) and technologies (*in silico* modelling systems, repositories of standard biological parts) [34, 35]. This engineering approach featuring a model-based rational-design, was first proven successful with the publication of the first genetic switches, the repressilator and the toggle switch [36, 37].

These have laid the foundation of a promising field whose potential applications fostered the creation of dedicated programmes advancing its key enabling technologies and expanding its applications by creating a well-knit community,

yielding remarkable breakthroughs and potentiating our ability to engineering biological systems. The ‘tipping point’ for broad, market-meaningful adoption of Synthetic Biology came with the arrival of dramatically cheaper high-throughput DNA synthesis and sequencing, easily-employed biodesign tools and the availability of public repositories (Figure 3) [38]. The rapid adoption of these technologies by the expanding Synthetic Biology community provided evidence of a growing market, encouraging competition and further innovation targeting the creation of user-friendly toolkits and services accessible for all kinds of end-users. Consequently, the scope for synthetic biology has transcended from an emerging discipline to a foundational technological framework adopted widely in research and industry [39].

Now, Synthetic Biology is applicable to many areas; general bioengineering, editing of genomes of organisms in order to improve human health, transforming microorganisms to factories for producing certain drugs, creating cell-free systems capable of mimicking a cell’s machinery or constructing unnatural molecular biology with non-canonical molecules and interactions to be used in diagnostics [33]. The engineering potential for bacteria using Synthetic Biology is immense and innovations are almost limitless. With Synthetic Biology, it is possible to transform bacteria into production vehicles for biomolecules, to design biomolecules to our specifications, and to control the behaviour of the vehicle and the biomolecule production. For example, we can exploit bacteria as biochemical factories by creating new enzymes to produce desired chemicals [40-43]; bacterial genomes can be edited to render the chassis-cell compatible with a given strategy [44]; the cell’s environmental sensing may be influenced, and much more [45]. Synthetic Biology is now finally delivering the early promise of bacteria & cancer therapy.

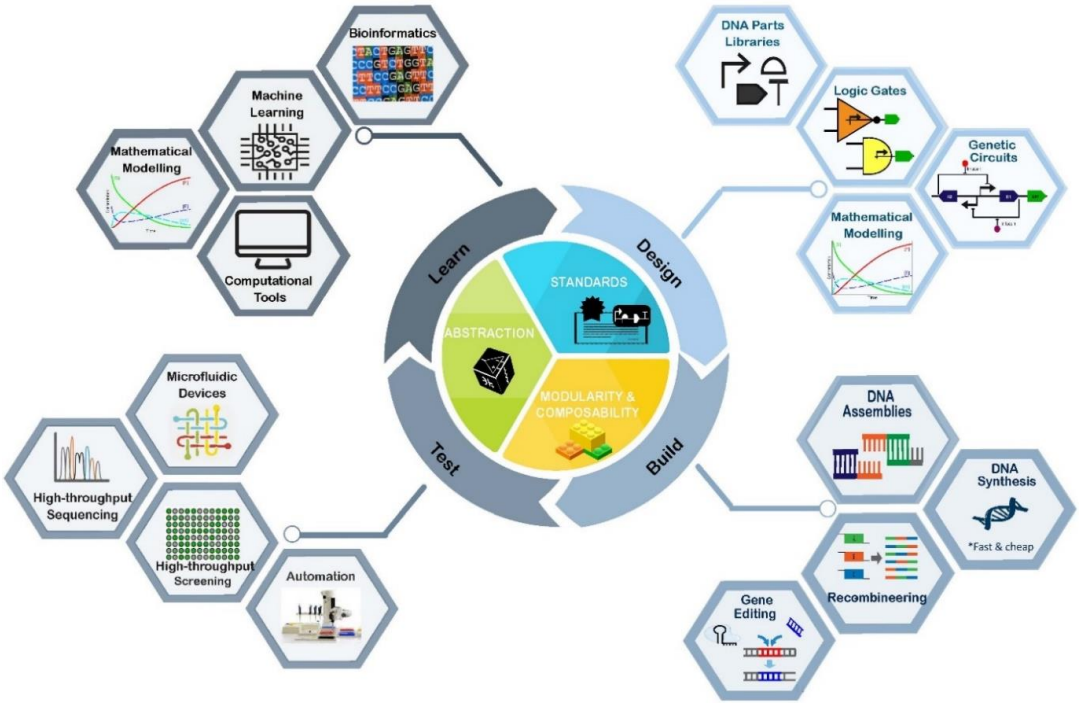


Figure 3. Synthetic Biology's design, build, test & learn (DBTL) cycle.

The foundation of synthetic biology lies in the introduction of engineering principles that enables the DBTL cycle.[35] In this figure are also portrayed the different technologies developed by the synthetic biology community for the advance of the DBTL cycle [46, 47].

Path to market

Full engagement with the Synthetic Biology approach goes beyond the scientific aspects of a technology, and incorporates all stages of R&D required to achieve an appropriate product. The SB process embraces, from the idea stage, multiple actors/stakeholders along the product development chain. The Design-Build-Test approach (see later) and rapid prototyping capacity of Synthetic Biology facilitates incorporation of design/redesign input to address multiple needs, at earlier, cheaper stages of R&D, before it is too late. The power to bestow sophisticated properties on bacterial chassis, devices and biomolecules permits early addressing/pre-empting of aspects of safety, efficacy in the field, scale-up etc., in addition to reducing the duration of the product development path for a product, thereby cost & risk of medicine development and therefore the final cost of the actual product (Figure 4).

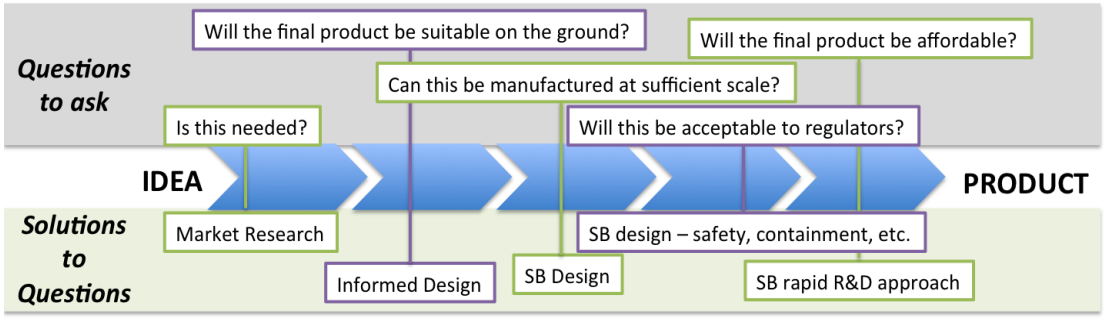


Figure 4. Synthetic Biology 'Built-In' Market-driven R&D Considerations *(SB – Synthetic Biology)*

The synthetic biology approach to bacterial engineering

Synthetic biology borrows ideas, concepts and lingo from the engineering world and applies them to biology. In nature, complex systems comprise highly interconnected entities performing synchronized functions. However, synthetic biology, applies engineering principles (modularity, composability, abstraction, and standardisation) to redefine them into a modular and composable way. Through this framework, the elementary unit of a system is a thoroughly characterised and standardised 'part' – a motif (DNA sequence or genetically encoded product) with a defined task in a coding region. These motifs are the building blocks of a 'Lego like' scheme, where they are mix-matched to build fully functional genetic 'devices', capable of performing a defined function and an established input/out relationship. Devices are integrated into a chassis (e.g. a bacterial cell), to build a 'system', capable of producing a targeted biomolecule or behaviour (Figures 3, 5) [35, 48-50].

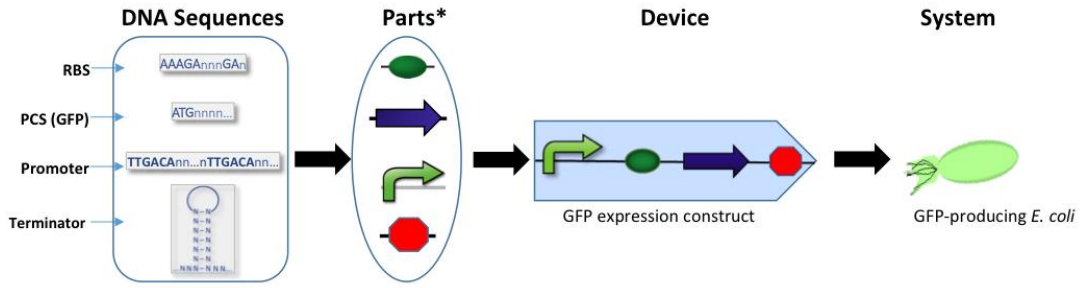


Figure 5. Schematic representation of an abstraction hierarchy. *Here, a genetic component, (a gene, transcription factor or a promoter) is defined as a 'part'; a collection of parts that together have a defined function = a 'device'; a collection of devices integrate to create 'systems'. (RBS: Ribosome binding site; PCS: Protein coding sequence).*

Furthermore, in order to achieve a logical form of cellular control through rational design, synthetic biologists apply electrical circuit analogies to describe genetic networks and biological pathways. In this context, a 'circuit' is a network-like composition of parts and/or devices, perform logical operations, that can be modelled, e.g. 'if' X condition is met, 'then' provide Y output [45].

Advancing the design-build-test cycle

The expansion of open-access catalogues of thoroughly characterised biological parts in computer readable formats, has advanced the rational *design* of biological systems [51]. Advances increasing our capabilities for DNA synthesis and assembly [7], and genome-scale engineering [52], and their translation into automated, high-throughput systems have potentiated our *building* capabilities, and increased their standardisation, efficiency and reproducibility.

The thorough characterisation and measurement of a system's functionality *(test)* in 'real-time', is now made possible through high-throughput quantitative analysis tools that provide feed-back, facilitating the parameterisation of predictive models (See Figure 3) [51]. Altogether, these advances accelerated the pace of *design-build-test* cycle, and allowed the construction of highly sophisticated systems, built from multiple components and implying multiple layers of cellular regulation [53]. The arrival of systems with higher complexity, brought along a new level in the abstraction hierarchy: biological 'modules'. These are subsystems made from a collection of discrete and defined devices with interconnected functions that together perform a complex task, as part of a higher wholesome system. Such operate as pathways resembling integrated circuits [48, 49, 54].

In this context, intelligent and tuneable systems or circuits, are made possible by integrating parts with a thoroughly characterised function. Parts catalogues, now supply a vast number of parts (sensors, regulators, actuators). These are constantly enriched with *de novo* parts harvested from nature, or *variants* created by predictive modelling (iterative rational design) or directed evolution [50]. Expansion that paved the way for the creation of regulatory elements (devices, modules) capable of

manipulating different biological processes, simultaneously. Beyond transcription, synthetic systems now include modules regulating translation [55], post-translational modifications [56], and epigenomics [57, 58]. Novel parts advancing a multi-layered control include: CRISPRi [59], recombinases [60] invertases [61] feed-back and feed-forward loops [48, 62-64] for transcription; ribozymes and riboregulators [65-67] for post-transcriptional processes; and novel receptors [68], secretion tags, degradation tags, protein-binding tags for post-translational processes [50].

These provided the building blocks for building regulatory devices with logic behaviour, such as: switches [60, 69], logic gates [70, 71], stable oscillators [72], Riboswitches [73], and diverted scaffolds [74-77]. Similarly, these devices have now been applied to develop systems integrating logic to create permanent memory or produce complex calculations [78, 79], wire circuits through quorum-sensing [80, 81], building genetic edge detection programmes [82], controlling multicellular migration pattern and population growth [83], and building layered logic programmes enabling the construction of large integrated circuits in a cell [78]. There is an abundance of literature demonstrating the diversity and potential of these systems [45, 51, 84].

Applying synthetic biology principles for i*n situ* biomolecule production by bacteria now offers controllable strategies to externally controlled or self-regulated (intelligent) chassis cell and device behaviour (see later). Since much of the foundational work on Synthetic Biology was carried out on microbes including *E. coli*, the technical knowhow for sophisticated modifications for heterologous agent production, controlled expression, and safety-attenuation is readily available for deployment in the setting of *in situ* therapeutic production [85]. Synthetic Biology can improve this technology at all levels; i) the vehicle; ii) the production of the biomolecule carried by the vehicle; and iii) the biomolecule's activity.

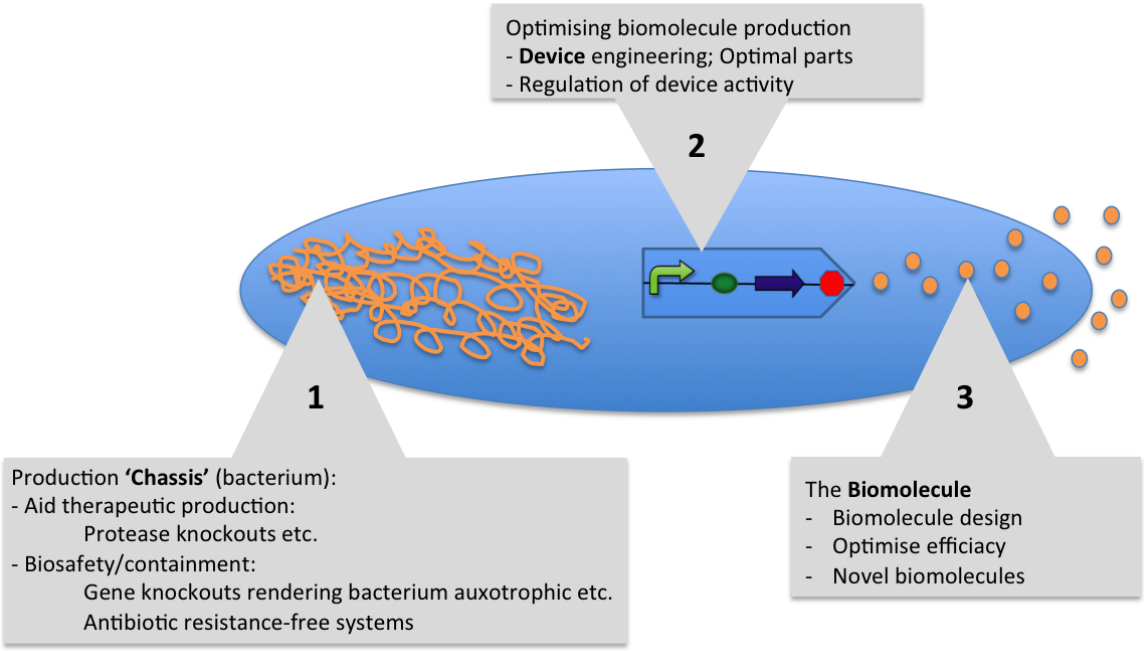


Figure 6. Synthetic Biology improves the technology at all levels.

1. *The chassis cell (through bacterial genome engineering);* ***2.*** *The production of the biomolecule by the system (through device engineering (including regulation of device activity));* ***3.*** *The biomolecule (e.g. modelling to obtain the optimal final biomolecule).*

Cancer as an example indication

In the cancer context, bacteria are being investigated for biomolecule production/delivery both locally (direct therapy), and distally to tumours (within the GIT; immunotherapy) [27, 86-88].

Bacterial growth in tumours

Various studies have shown that tumours support the growth of different bacterial species. A tumour microbiome has been described by different laboratories [89-92]. Separately, both in clinical and pre-clinical studies, different bacteria have been shown to preferentially colonise and proliferate within tumours following systemic administration [86, 88, 93]. It is believed that bacteria in the bloodstream leak from the abnormal vasculature within tumours and lodge locally where they are protected from the immune system due to the immune-suppressed microenvironment of tumours. The 'targeting' process therefore is more of a passive phenomenon of

selective growth, without the involvement of chemo-attractants and relates to the tumour environment being permissive to bacterial survival and replication, unlike most healthy tissue. Chemotaxis may play a role post tumour targeting, influencing the manner in which certain bacteria distribute within the tumour [94]. Further parameters that distinguish tumour from healthy tissue include nutrient availability to bacteria (from tumour cell turnover in necrotic regions) and regions of low oxygen potential (where anaerobes and facultative anaerobes can grow optimally) [87, 93].

Bacterial tumour-targeting technology is based on the bacterium to selectively survive and replicate within solid tumours, growing to high concentrations, where they can 'pump out' therapeutics or locally activate agents. Depending on the strategy, the bacterium itself (the chassis) may possess intrinsic oncolytic properties (often the case with pathogens), or may have no effect on tumour growth unless engineered to produce an agent. This platform technology is applicable to a wide range of therapeutic or diagnostic strategies. Clinical trials have demonstrated the safe use of live engineered bacteria in cancer patients, and preclinical studies using modified bacteria as tumour-selective agents have demonstrated the high potential for bacterial-mediated cancer therapy via *in situ* biomolecule production [29, 86, 93].

Bacteria in breast tissue

The residency of bacteria in breast tissue has been affirmed in several studies documenting the microbiota of healthy breast tissue, mammary glands and breast milk [91, 95-100]. As a whole, the breast is a favourable environment for the growth of bacteria, as it is made up of fatty tissue, with extensive vasculature and lymphatic drainage [101, 102].

Studies to determine the diversity of bacterial species found in the breast suggest that there is a more diverse array of species compared to many other body sites [103]. Interestingly, these bacteria have important roles attributed to them in supporting the healthy development and immune maturation of neonates [101, 104, 105]. The breast microbiome has been suggested to be derived primarily from the microbiota of the overlying skin and the oral microbiome [95, 98, 100]. This facilitated by ductal

openings at the surface of the nipple that allow their entrance from the environment, skin and/or mouth [106].

Sampling of the microbiome supports this colonisation as it has been reported that the breast microbiome is quite similar in composition to that of the skin [95, 98, 100], but that this composition shifts towards the oral cavity once breastfeeding begins [99, 101]. However, the involvement of the gut as a source of these bacteria cannot be ruled out. During the late stage of pregnancy and lactation, physiological changes occur which allows for an increase in bacterial translocation in the gut. This is facilitated by dendritic cells, which cross tight junctions in the gut epithelium and transport bacteria from the gut lumen to the mammary glands [101, 104, 105]. It is plausible that certain bacterial species inhabiting the breast have health benefits beyond those conferred during lactation.

A healthy microbiome can deter the invasion and growth of pathogens and provide protective immune stimulation against disease. Certain bacterial strains found in the breast have been shown to produce lantibiotics, a class of bacteriocins capable of limiting the growth of pathogenic bacteria which could trigger chronic inflammation leading to malignancy if otherwise left to proliferate unchecked [97, 100]. For instance, the oral administration of certain lactobacilli has been shown to be effective in preventing and treating mastitis in women [107]. Indeed, the production of milk oligosaccharides influences the breast microbiome and is key to the establishment of the infant gut microbiota [108]. Moreover, several probiotic bacteria can modulate the immune system to supress inflammation or may serve to trigger an antitumour immune response [109-111].

Certain bacterial species found in the breast tissue, such as *Lactococcus lactis*, have also been shown to increase the expression of anti-inflammatory response pathways or activate natural killer cells capable of controlling tumour growth [91, 97, 109, 112, 113]. In fact, epidemiological studies have found a strong correlation between the consumption of fermented products and a reduced risk for breast cancer [102].

Another interaction with host cell physiology, which may play a protective role against the development of cancer is the metabolism of oestrogen and phytoestrogens [114]. The microbiota, namely species such as *Clostridium and Escherichia* can increase

circulating levels of oestrogen via deconjugation of sulphonated oestrogens by β-glucuronidase, thereby associating these bacterial strains with an increased risk of breast cancer [115]. Conversely, the metabolism of dietary phytoestrogens into bioactive molecules such as equol, urolithins and enterolactone, that compete with human oestrogen at its receptors and can thus reduce oestrogen-driven breast neoplasia [116, 117].

Moreover, bacteria found in the breast have been associated with the production of antioxidants that neutralise free radicals [95, 97]. Supporting all the aforementioned benefits to breast health provided by bacteria, is the evidence raised by large clinical studies correlating the use of antibiotics with an increased risk of breast cancer [118].

The role of the microbiome in immunity

Numerous studies have now provided evidence of the pivotal role of the gut microbiome in the development and regulation of our immune system, both locally and systemically. Locally, the gut microbiome has been found to promote the development and maintenance of a robust mucosal layer and associated epithelial and lymphoid tissue, where it regulates the maturation of dendritic cells (DC) that activate naïve T-cells and enhances the secretion of Immunoglobulin A. This effect has been recently found to include all mucosal tissue in the body where it regulates adaptive immunity [119, 120].

Systemically, the gut microbiota has been shown to modulate both the innate and adaptive immune system. For the innate immune system, microbial molecules (LPS, SFCA, and Peptidoglycans) have been found to stimulate myelopoiesis/granulopoiesis of granulocyte-macrophage progenitor cells ensuring homeostatic levels of neutrophils, monocytes and macrophages. In addition, certain microbial antigens, such as SCFA, have been found to increase dendritic cell differentiation by enhancing haematopoiesis of their precursors [121]. [37]. The microbiome has also been found to regulate the adaptive immune system by promoting the proper development of distal lymphoid tissue that harbour the development of B- and T- cells [122]. The adaptive immune system is also regulated by the APC nature of DC, which present microbial antigens to B- and T-cells triggering their differentiation. Importantly, this mechanism

has been found essential for the pro-/anti- inflammatory balance of the immune response. In this regard, the microbial composition of the microbiome can influence the CD4+ Th1 to Th2, and the Th17/Treg balance, where it has been shown that by introducing bacterial taxons derived from high fibre diets, skewed ratios of pro-inflammatory cytokines can be alleviated and immune balance restored [119, 123, 124].

The role of the microbiome in cancer immunotherapies

The intricate connections between our immune system and our microbiome have been found to influence the outcomes of immunotherapies. This was first found for immune-check point inhibitors, namely, anti-PD1/PD-L1 and anti-CTL4 [122, 125-127]. Clinical studies reported a higher response to therapy and overall survival in patients with diverse microbiome profiles (eubiosis), containing Bifidobacterium, Akkermansia and Enterococcus, among others. On the other hand, patients with a reduced microbiome diversity (dysbiosis) responded poorly to this therapy [128, 129]. The effects of these bacterial profiles were confirmed in GF murine models where mice that received FMT from responders also exhibited an improved response and those who received an FMT from non-responders developed resistance to the therapy that was reversed upon administration of Akkermansia and Enterococcus. It was observed here that responders had an increased immune infiltration in the tumour microenvironment with a marked increase CCR9+CXCR3+CD4+ and $CD8^+$ T-cells, and a decrease in Treg cells [130, 131]

Similar results have been shown in murine models for anti-CTL4 (ipilimumab), where the efficacy of the therapy was shown to be reliant on the microbiome. Here, ablation of the microbiome reduced response to treatment and introduction of taxons corresponding to the Bacteroides and Burkholderia genus to non-responders triggered a Th1 response and promoted DC maturation, restoring the efficacy of the therapy. Interestingly, some studies have shown that the therapy can induce detrimental changes to the microbiome that may influence the development of resistance [132]. This highlights the relevance of managing the microbiome richness during the course of treatment [127]. Murine studies with other immunotherapy agents, namely anti-IL-

10 and Adoptive Cell Therapy, have also shown that the microbiome stimulates the effectiveness of these therapies [122, 127].

Non-tumour targets for cancer therapy

The vaccination using live microbes field is, by comparison to the above, a mature area of research with significant commercial interest that employs different types of microbial vehicles including modified viruses or bacteria, which confer immunological responses against infectious diseases or cancer. The goal of cancer vaccines is to break tolerance of the immune system to specific antigens known to be expressed mainly or exclusively by particular tumour cells - tumour-associated antigens (TAA). Bacteria are advantageous as antigen delivery vehicles due to their ease of bioengineering and diverse collateral effects on the immune system.

As part of their natural life cycle, infectious bacteria, following entry to the body, are internalized by phagocytes, followed by MHC presentation of their antigens to the rest of the immune system. Through addition of synthetic antigens to a bacterial system, the process can be hijacked to mount a host immune response to a desired antigen (e.g. tumour-associated). Used in this setting, the chassis delivers an antigen to antigen presenting cells (APC), such as M cells in the gut mucosa, and does not involve growth in tumours. The bacterium is safety attenuated to render it non-infectious, and equipped with a device to produce specific tumour antigens (either genes or proteins).

The vehicle itself also induces a desirable immune response in the vaccine context (similar to an adjuvant). Following administration (per oral, intramuscular, or intravenous), the bacterium is taken up by the patient's antigen presenting cells. The bacterium releases genetic material or antigens into the immune cells that then initiate a systemic immune response specific to the target antigen.

There are multiple safety-attenuated strains under study as vehicles for vaccination; *Listeria monocytogenes, Salmonella, E. coli* and strains of *Shigella, Lactobacillus* and *Yersinia* of which *L. monocytogenes* (Lm) and *Salmonella* are being studied clinically [133, 134]. Significant, highly promising therapeutic outcomes are being realised from

these vaccine platforms in multiple Phase II and III trials with patients of disparate cancer indications [i] [134].

Synthetic biology approaches to improvement of bacterial agents and treatment strategies

The chassis cell

Safety attenuation. Employment of bacterial strains with a natural ability to survive and grow within human tissues (i.e. pathogens) is attractive from an efficacy standpoint, but obviously undesirable from a safety perspective due to off-target growth within healthy organs, coupled with recognition by the patient's immune system as a disease-causing agent. Strain attenuation can be used to limit capacity to survive in non-target healthy tissues e.g. liver, or to reduce pro-inflammatory reactions.

Traditionally, attenuation was achieved by random mutagenesis of a wild type strain and selection for certain favourable phenotypes e.g. tumour invasion, proliferation etc. Purpose-designed systems are preferable, involving editing of genes that are known to be involved in pathogenesis. For example, *msbB* and *purI* are two genes that have been eliminated from the genome of *Salmonella* in order to create VNP2009 [135] the first *Salmonella* clinical trial agent. Another attenuated *S.* Typhimurium defective in guanosine 5′-diphosphate-3′-diphosphate (ppGpp) synthesis (a molecule responsible for regulating salmonella pathogenesis [136]) was also generated by genomic editing. Similar editing can also reduce unwanted host responses to non-pathogenic bacteria; e.g. the probiotic *E. coli* Nissle 1917, which is part of our natural gut microbiome, has been attenuated via an *msbB* deletion which reduced pro-inflammatory cytokine stimulation compared with wild type [137].

Cell targeting. Although bacteria do not actively home to tumours, it is possible to improve their specificity to the tumour environment and limit their ability to proliferate in healthy tissue through exploitation of unique tumour traits to guide

the design of more tumour-selective bacteria. For example, Yu *et al* [138] restricted the growth of bacteria to hypoxic regions, a phenotype found only within tumours inside the body. An essential gene for cell wall synthesis, *asd*, was placed under a hypoxia-inducible promoter (*PpepT*) which allowed expression to take place only under hypoxic or anoxic conditions. In parallel, a second device expressed the antisense of *asd* under an aerobic promoter (*PsodA*). This device inhibited growth under normoxic conditions. Integrating both devices into a module, enabled a logic gate restricting replication to areas with low oxygen concentration, such as those found inside the tumour. Such a circuit would eliminate the capacity of bacteria to grow in healthy tissue, thus adding another layer of safety. 'Trapping' bacteria within tumours can also be achieved via addition of tumour cell ligands. Using a sophisticated surface display system, the peptide RGD was surface tethered to *Salmonella* in order to improve its targeting capabilities towards specific integrin expressing cancer cells [139]. Similar strategies could be used in other systems to target bacteria to specific cells/tissues. Such levels of bioengineering sophistication can upgrade chassis cells in both efficacy and safety.

Bacterial vs Viral chassis. Both, bacteria and viruses are effective delivery vehicles for different cargoes. Here we outline the characteristics that will determine their feasibility under different scenarios.

Table 2. Bacteria v Viral vectors.

Pro-Bacterium	Pro-Viral vector
Bacterial chassis = final biofactory. Multiple components of the biofactory cell genome can be engineered *in vitro (see Figure 4 part 1)*	
Bacteria can generally carry more devices	
Bacteria can produce biomolecules independent of / external to host cells.	
	If biomolecule must be delivered internally to host cell, viral vector transduction efficiency is much higher than bactofection
Viral vectors must be invasive => safety concerns.	
	Viral vector better as *in situ* host cell (biofactory) editor.
Antibiotic sensitivity can act as safety 'Off switch'	

Bacterial manufacture cheaper	
Bacterial biomolecule type may be nucleic acid, protein or small molecule, while viral vector biomolecules are restricted to nucleic acid	
	Bacterial expression of eukaryotic gene sequences may not be as efficient as with viral vectors.
Bacteria may naturally colonise and replicate in specific tissue/location, more so than viruses.	
Bacteria more transient than viral vector (safer is some circumstances)	Bacteria more transient than viral vector (viral vector better for integrating device in host cell genome)
	Viral vectors have a closer relationship to human cells.

Biomolecule delivery and production

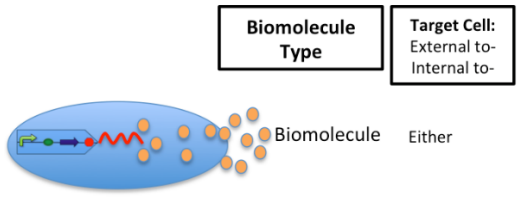
There are two broad ways to deliver a biomolecule in the bacterial context – i) at the tissue level, normally external to target cells, or ii) internal to target cells. The delivery modality must be matched with the biomolecule's therapeutic modality. For several therapeutic strategies, simply 'flooding' the environment with bacterial-produced protein is sufficient, and non-invasive chassis are suitable, and from a safety perspective, desirable.

Bactofection. Delivery of biomolecule internal to cells involves chassis lysis after which its contents are released to the cytoplasm of the target cell. In this context, the biomolecule may be protein, RNA or DNA depending on the strategy employed. This strategy is often referred to as bactofection (bacterial transfection).

Bactofection can be 'active', involving an invasive bacterium mediating its entry to a cell, or 'passive', as is the case with phagocytic immune cells [140]. 'Smart' target cell entry may be achieved through Synthetic Biology approaches, using devices that sense different inputs leading to an invasive output. Host cell invasion by *E. coli* was achieved by expressing the protein *inv* gene from *Yersinia pseudotuberculosis* which was triggered by hypoxia, cell density or an exogenous inducer [141]. Once the bacterial cells came into proximity with the host cell membrane and reached a certain density, the circuit became activated leading to the production of *inv* gene resulting in

tumour cell invasion. Some strategies utilise occurrences post-invasion, for example van Pijkeren *et al* [142] devised a system by which a lysin was expressed only following host cell internalisation, in order to induce a cascade of bacterial lysis.

Types of biomolecule 'payloads' and optimal production. There is a large and diverse collection of biomolecules which have been investigated in studies with bacteria to date, and may be peptide-based, RNA or DNA in nature (Figure 7).



Biomolecule Type	Target Cell: External to- Internal to-
Biomolecule	Either
RNA	Internal to
DNA	Internal to

Examples

Delivery Modality	Biomolecule example	Target cells	Strategy
Local Secretion	Peptide	Host cells	Anti-inflammatory
	Small Molecule	Cancer cells	Toxicity Induction
		Immune cells	Vaccination
		Local microbes	Anti-Pathogen

Delivery Modality	Biomolecule example	Target cells	Strategy
Bactofection	miR	Phagocytic cells	Cell Programming
	shRNA		

Delivery Modality	Biomolecule example	Target cells	Strategy
Bactofection	Plasmid DNA	Cancer cells	Toxicity induction
		Phagocytic cell	Cell Programming

Figure 7. Example types of biomolecules and relevant medical strategies

Controllable and intelligent systems. Currently, a tight regulation of multi-module circuits is more easily achieved, increasing the predictability of desired phenotypes. This applies to biomolecule production, which kinetics benefit from a sophisticated control in gene expression. By applying a rational design, different layers of control over biomolecule production and/or the vehicle can be applied, when appropriate for a chosen strategy.

Such a system can incorporate a sensing module able to respond to numerous stimuli. A rich-repertoire of now available, characterised sensory parts, enables the creation of systems capable of responding to a variety of physical or chemical inputs, such as oxygen concentrations, acidity, cell density, drugs, molecules, radiation. Sensors are often built upon promoters whose activity can be regulated by environmentally

responsive DNA binding protein [50, 143, 144]. Here, regulation is mediated by the binding of the protein into a transcription activator or repressor site in the promoter sequence. This generates a conditional 'switch (ON/OFF)' behaviour – nominated positive or negative feedback loops. A combination of these parts can be used to create AND/OR/NOR gates [45]. Depending on requirements, designs can range in flexibility and sophistication. The design of complex systems, whose multi-components' interaction rely on multiple factors, (e.g. DNA-protein binding/dissociation constants, kinetics and other biophysics), is now made possible by *in silico* analysis. These tools enable the prediction of such level of regulation by applying mathematical based, known biophysical constants and coefficients to model biological processes [145, 146].

An early example of a controllable system in this context involved an engineered *Clostridium* [147]. These authors created a switch turned on by radiation that could trigger the production of a protein with therapeutic properties (e.g. TNFα, cytosine deaminase (CD)) and induce a cytotoxic response in preclinical models. In an analogous manner [148] used a device switched on by the sugar arabinose to give a toxic output in order to treat colon carcinoma.

More recently, circuitry was taken to the next level. A circuit was designed whose input is cell density but leads to several outputs regulated by a common part. The circuit is composed of an activator, a reporter, a therapeutic and a therapeutic gene delivery device [149]. The circuit is based on the quorum sensing system *lux*. *LuxI* catalyses the synthesis of N-3-oxohexanoyl-L-homoserine lactone (AHL) which freely diffuses and accumulates in the surrounding proportional to cell density. AHL activates the transcriptional activator *LuxR* to activate genes that have a downstream *luxI* promoter. The *LuxI* promoter itself was inserted in front of *luxI* gene in order to create a positive feedback regulation to support the integrity of the circuit. GFP was used to give a light signal output. The bacteriophage lysis gene (ϕX174 E) was used to aid bacterial lysis and deliver the cytotoxic payload, and finally, the payload itself was the cytolysin, a pore forming protein. In an analogous circuit, two parallel devices where employed to deliver a cytotoxic payload to tumours in mice [150]. Therapeutic protein production was controlled by salicylate and lysis of bacterial cells was controlled by tetracycline. Such a system first allows bacteria to target to tumours

without putting a metabolic burden on them. Production begins only after bacteria reach optimum numbers within the tumour, and lysis serves to deliver the therapeutic protein to the surroundings in the most efficient manner. More recently, the Hasty group engineered an elegant 'synchronized lysis circuit' in *S.* Typhimurium to induce lysis at a threshold population density (through quorum sensing) and release its therapeutic cargo [81].

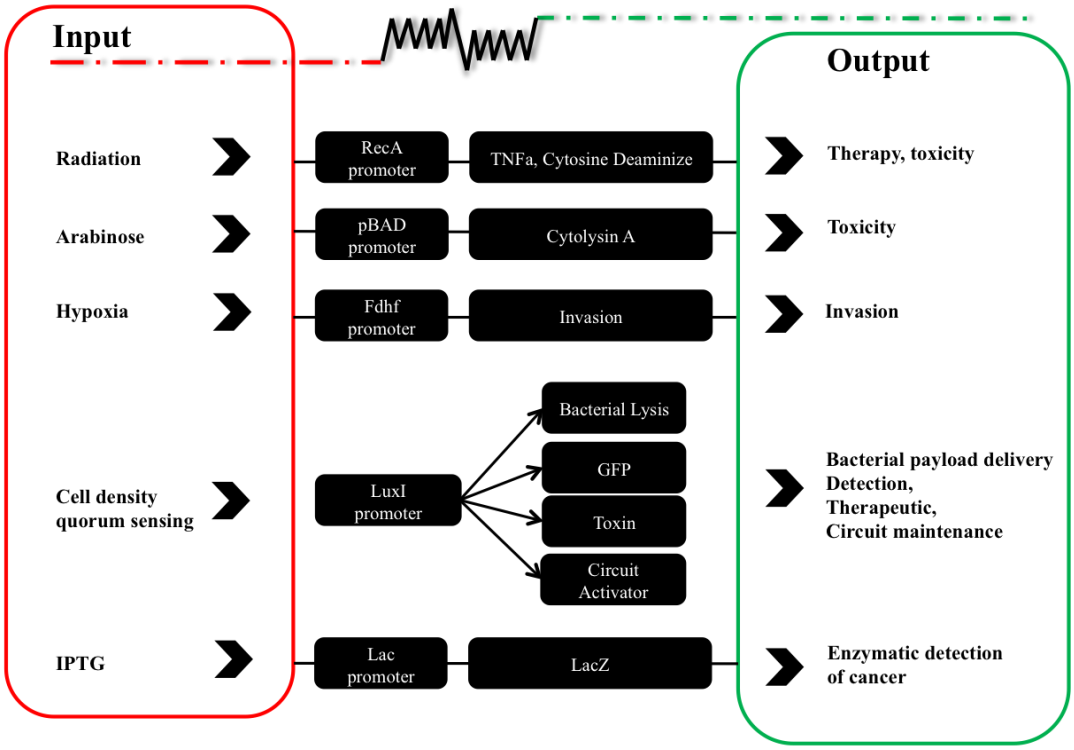


Figure 8. Examples of controllable/intelligent bacterial systems in oncology studies

Treatment strategies

As Synthetic Biology became more sophisticated, new possibilities became realized. Strategies could now be re-designed to deliver maximum efficacy. We now have the capacity to deliver protein, RNA, DNA and to activate small drug molecules specifically at bacterial-specified sites. Production of biomolecules can now occur in bacteria and/or host cells and parameters such as the kinetics, location and level of production and the function of the product itself can be controlled.

Therapeutic Production

Peptide production. Bacterial production of peptides is well described in several research domains, and is commonly used industrially for recombinant protein production [151, 152]. In *in vivo* strategies, non-invasive/apathogenic bacteria are primarily employed in this context. As an example, Lactic Acid Bacterial chassis producing *in situ* antiproteases and antioxidant enzymes have been tested successfully for their prophylactic and therapeutic effects in murine models of colitis [153]. Numerous intratumoural bacterial production of various peptides at both clinical and pre-clinical stages have been described [29, 154]. A wealth of technology has been developed for optimisation of protein production by bacteria [155].

Bacterial cells are enveloped by sophisticated membranes that regulate what enters and exits the cell. In Gram-negative bacteria, for example, the cytoplasmic interior is separated from the exterior by a thick outer membrane, a periplasmic space and an inner membrane. Depending on the nature of the therapeutic biomolecule used, cytoplasmic expression may hinder its activity. A number of systems exist that place biomolecules of interest in different compartments of the bacterial cell or secrete them to the exterior. Secretion to the surrounding environment is frequently desirable, and a number of systems are available for different bacterial genera employing signal sequence 'parts' in devices to promote appropriate secretion [155, 156]. Surface display parts can direct proteins to the outer membrane of bacteria [139, 157]. Recombinant proteins commonly surfaced exposed are antigens and antibodies [158].

RNA production. Small interfering RNA and microRNA has generated much interest in recent years in both basic and applied biology. For example, *S.* Typhimurium has been utilised in various preclinical cancer studies as a chassis to deliver small hairpin RNA (shRNA) against GFP, STAT3 or bcl-2 [159, 160].

Small molecule activation. In order to address the problem of target specificity of small drug chemotherapy, researchers have been using synthetic biology to enable bacterial-colonised tumours to act as the final stage of toxic drug 'synthesis'. Here, enzymes are produced by bacteria at the tumour site, while the chemical reactants (prodrugs) are administered later. The active drug generation takes place at the tumour site, mediated by the bacteria which enzymatically activate the actual

chemotherapeutic (reviewed in [161]). Recent work suggests that multiple drugs can be activated concurrently [162] and opens doors to new ideas such as having devices that can concurrently activate *in situ,* multiple drugs with diverse mechanisms of action in order to overcome drug resistance.

Host cell modification

In some circumstances, it may be desirable to induce the host cell to produce the biomolecule itself. Invasive chassis deliver devices or biomolecules to mammalian cells by bactofection. Such devices feature parts that are compatible with eukaryotic environments, and therefore switched on post-delivery. Usually, the specificity in such systems comes from bacteria and the devices themselves have a constitutively active switch (a eukaryotic promoter such as CMV) which fires upon delivery to any host cell. However in other cases, another layer of regulation can be introduced at the device itself, by using a switch that is only turned on by cancer cells (though use of a tumour-selective promoter [163]) therefore providing an extra level of specificity as well as therapeutic potency.

Delivery of eukaryotic devices is not limited to cancer cells, or invasive bacteria. Byrne *et al* used a non-invasive *E. coli* to infect phagocytic cells (Tumour Associated Macrophages (TAM)) and deliver DNA modules that produced light as an output [140]. In this case, the specificity towards the phagocytes was brought by the 'non-invasiveness' of the bacteria. Vaccine strategies employ a similar strategy.

Diagnostics

Co-localisation of a bacterial agent with a specific site/cell type presents opportunities for diagnostic strategies. For example, in the context of oncology, tumour detection can either be direct, for example by intratumoural bacterial imaging, or indirect by biological fluid analysis of biomarkers (liquid biopsy). In this context, representing a prototype Point of Care test, a prototype system has been developed to detect cancer by urine sampling. *E. coli* expressing a regulated *LacZ* was constructed in order to

detect murine liver tumours [164]. Following bacterial colonization of hepatic tumours in mice, bacteria express *LacZ* enzyme following induction by IPTG. Subsequently, a derivative of luciferin is administered which is cleaved by *LacZ* to pure luciferin and cleared through the urine. Luciferin is then measured by emission of light directly from the urine sample offering quick non-invasive tumour detection. Similar to the above, [165] created an inducible reporter/biomarker module that can be detected in blood samples by antibodies in an ELISA type assay. The biomarker, ZsGreen expressed by *Salmonella,* was shown to be suitable for detection of colon carcinoma in mice.

Regulatory agency aspects

The expanding scope for and adoption of biological engineering applications potentially presents the need for our current ethics and governance to evolve also [38]. If Synthetic Biology actors approach this aspect correctly, regulatory concerns can be overcome. Currently, Bacillus Calmette-Guerin (BCG), a local treatment for bladder cancer, is the only live bacterium in clinical use, and is not genetically modified. However, precedents for licencing of live GM microbes have been set, with viral-based chassis. 2012 yielded the first licensing of a gene-based therapy in the western world, with the EU EMA licencing of Glybera - an AAV chassis engineered to express lipoprotein lipase in the muscle of deficient patients [166]. Talimogene laherparepvec (also known as T-Vec) was approved by the FDA in 2015, with the brand name Imlygic, for the treatment of advanced inoperable melanoma. In 2016, it was approved in Europe. It is an oncolytic virus and consists of a genetically modified Herpes Simplex Virus (HSV) chassis carrying a device producing *in situ* a cytokine (GM-CSF) that helps to induce immune responses following intralesional injection. [167]

Engineered bacteria for vaccine use have advanced to late stage clinical trials and therefore the safety/regulatory aspects of live GM bacteria are also being tested concomitantly. Clinical candidates have medical and environmental safety requirements, which can only be met by the use of bioengineering, involving biological containment of both the vehicle and any ‘non-natural’ DNA elements. For

example, Aduro Biotech has been developing a *Listerial monocytogenes* agent for use in patients [168]. An attenuated form was created by deleting two genes critical to pathogenicity – *internalin* B and *act* A, while antigen gene cassettes are inserted in the bacterial genome therefore obviating antibiotic use [169]. To maximize agent production, it may be desirable to maintain the antigen gene on an episomal plasmid in order to increase gene copy number.

While plasmid maintenance in the lab environment employs antibiotic resistance modules, this is not acceptable for a market product from a regulatory aspect. Alternative plasmid maintenance systems have been created, based on modifications of both chassis and plasmid. Conditional or Balanced Lethal Systems involve genes required for bacterial survival being deleted from the genome of a chassis and transferred to a plasmid into which the device is also inserted. Bacteria produce the biomolecule as long as the plasmid is retained [170] and die in the event of a plasmid loss. There are many more examples demonstrating that Synthetic Biology offers realistic solutions for the development of bacterial systems in order to meet clinical requirements.

Concluding Remarks

Synthetic Biology is a burgeoning field that is driving the progression of bacterial agents in the health industry. The application of Synthetic Biology to improve bacterial agents for use in the strategies described is key to fulfilling earlier promises. Unlike before, intelligent precision engineering will permit the generation of effective agents. Further new developments pertaining to the regulation of bacterial safety will also be attractive to market stakeholders, paving the way for state of the art bacterial therapeutics. Perhaps the most valuable aspect overall, is the Synthetic Biology all-stakeholder-inclusive approach to R&D from idea to product. Thanks to Synthetic Biology, the time for developing successful bacterial-based disease treatments has finally arrived.

Online Resources

https://clinicaltrials.gov/ct2/show/NCT02853604

3. FFPE-induced DNA Damage; Relevance to microbiome analysis

As the fields of human genomics and human microbiome expand, so too does the need for access to high-quality nucleic acid samples which are truly representative of the genes/cells under study. Formalin-fixed, paraffin-embedded (FFPE) tissue specimens have now become a main source material for these studies. However, there is still missing a generalised consensus on the type and frequency of FFPE-induced sequence artefacts. A higher discerning capacity could be achieved by a better understanding of recent and foundational knowledge of formalin-biomolecule interaction and the effect that downstream treatments have on reacted molecules. Data quality could also be enriched by recently elucidated intrinsic DNA damage and mechanisms of repair. Use of FFPE samples is growing, and while most genomic research using FFPE samples has been focused on cancer, these samples are increasingly being exploited for other genomic research, such as microbiome surveys, where DNA quality has a larger impact.

A holistic understanding of FFPE-induced DNA damage, its impact on sequencing studies, and possible mechanisms for repair, stands to empower the overall genomics research community. Gathered experience and foundational knowledge enables design of suitable strategies for processing and repairing this sample type, as well as improving sequence analyses. In this review, we consolidate existing data on FFPE-induced DNA damage in human DNA. This includes a comprehensive review of the current state of understanding of formalin fixation and its interaction with DNA and other macromolecules, both *in vitro* and within the context of the native cellular milieu, as well as a discussion on the as yet uncharacterised FFPE-induced bacterial DNA damage and consequent effects on downstream microbiome analysis. Methods to repair FFPE-induced DNA damage to improve the quality of genomic analyses are discussed. Collating and considering this information highlights the value to be gained from increased focus on tools and approaches to addressing DNA damage inherent in FFPE samples.

Introduction

Formalin-fixed, paraffin-embedded (FFPE) tissue is the gold standard for pathology tissue storage. Tissue samples of this type represent the largest and longest time-spanning collections of patient material in pathology archives [171-173] and the availability of FFPE samples to be used as a source material for sequencing has been vital for progress in human genomics [174]. Numerous sequencing workflows enable use of these samples [175-180], which has in turn validated them as a viable and valuable source material for human genomics [181]. Investigations on the quality of DNA/RNA from human FFPE samples have revealed that processing and storage of these samples negatively impact the integrity of nucleic acids and the efficacy of their downstream analyses [182, 183].

As the sensitivity and specificity of sequencing strategies increased, attention was drawn to the study of the bacterial genetic material also within the body [184-186]. It is now well established that distinct microbial profiles can be found throughout the body, influencing human health [187-190]. As this field continues to expand, numerous body sites previously considered sterile have been found to harbour endogenous bacterial communities [191-196]. Such microbiome studies have traditionally availed of 'fresh' non-invasive samples (faeces, swabs etc.). However, microbiome studies targeting body sites for which sampling requires invasive procedures are constrained by access to samples [197, 198]. The use of FFPE tissue could open access to samples from cohorts of large numbers, accompanied by a clear medical/clinical history and thoroughly characterised histopathology report. These samples could be the source material for retrospective or longitudinal microbiome studies with sample numbers that guarantee statistical power. FFPE samples are already being utilised in microbiome research, albeit to a limited extent to date relative to genomics [91, 199-205].

In order for the potential value of increased usage of FFPE samples for microbiome research to be realised, the necessary workflows, protocols and quality control standards need to be in place [177, 206-208]. Microbiome studies to date have typically utilised approaches and tools designed for FFPE human sample analysis. Relevant to microbiome research, unique factors to consider in quality control of FFPE

samples are: 1) DNA fragment length suitable for 16S sequencing (460bp), 2) Presence of sequence alterations that may lead to false speciation events, 3) Low biomass – influence of contamination and host DNA, and 4) Unlike in human genomics where a single reference genome is available, microbiome genomes comprise DNA from a variety of bacterial, archaeal, fungal and viral sources. In many cases, the individual genomes themselves may be poorly characterised, or not at all. Accurate sequencing data is paramount in these circumstances.

Presented here is a thorough investigation of both foundational [209-218] and recent [177, 207, 219-225] research covering the interactions between formaldehyde and biological molecules to improve understanding of the mechanics of formalin fixation. Research using more sensitive techniques (MS, NMR) has provided new evidence on which molecules are more prone to these interactions, and their strength (*in vitro* and *in vivo*) leading to higher frequency. This also informs on the resulting type of DNA damage. Also discussed are the effects of downstream FFPE processing as additional sources of DNA damage.

Consolidating this information, coupled with new insights into intrinsic cellular DNA damage, can benefit sequence-based studies, such as cancer research where the aim is the detection of low frequency genetic variants present in a small number of cancerous cells [226], as they can provide a higher discerning capacity into the origins of sequence alterations. A considerable amount of NGS studies on FFPE samples have been published recently [177, 178, 206-208, 222, 227], revealing different patterns of sequencing artefacts, sometimes differing between similar studies. This lack of consensus could be explained by the nature of intrinsically damaged nucleotides and the molecular heterogeneity of samples analysed. In addition, after reviewing each type of DNA damage encountered in FFPE DNA, we propose pursuable strategies to repair such damage, representing the first review of repair strategies for FFPE-induced DNA damage. Finally, we highlight evidence of possible effects this might have in human microbiome sequencing studies.

The FFPE process

Standard histology processing of tissue samples begins with fixation in buffered formalin (4% formaldehyde in PBS) [228]. This tissue is then dehydrated with

increasing gradients of ethanol and cleared with solvents such as isopropanol and/or xylene and embedded in paraffin. Formaldehyde (HCOH) is a small electrophilic molecule, easily targeted by relatively strong nucleophiles, such as amino groups, generating crosslinks between them. In brief, HCOH crosslinking is initiated by a lone pair of electrons from a nucleophilic group attacking the partially positively charged carbon of HCOH, generating a methylol adduct in fast dynamic equilibrium. Upon dehydration, methylol adducts are converted into Schiff bases that can be stabilised into methylene bridges when these interact with other nucleophilic groups. (See Figure 1) [229, 230].

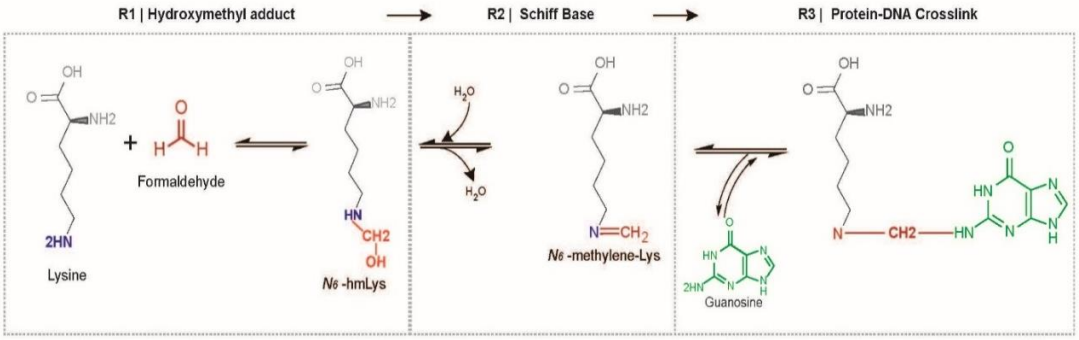


Figure 1. Chemical reactions for formation of formaldehyde derived methylene bridges between biomolecules.

In red: Formaldehyde. In blue – reactive amino group in Lysine. In green: dG. The reaction takes place in 3 steps: ***1)*** *Formaldehydes reacts with an amino group in an amino acid (most frequently).* ***2)*** *Upon dehydration, a reactive Schiff base is formed.* ***3)*** *Reaction of the Schiff base with a nearby reactive species, forms a methylene bridge (crosslink).*

Formaldehyde interaction with nucleotides

The chemical reaction of HCOH with nucleotides has been thoroughly characterised mathematically and experimentally *in vitro*. HCOH reacts with the endocyclic amino groups (NH) of deoxythymine monophosphate (dT) and the exocyclic amino groups (NH_2) of deoxyadenine monophosphate (dA), deoxycytosine monophosphate (dC) and deoxyguanosine monophosphate (dG), leading to the formation of hydroxymethyl adducts (hmN) in N_3 of dT (3-hmT), N_6 of dA (6-hmA), N_4 of dC (4-hmC), N_2 of dG (2-hmG) (Figure 2) [211, 213, 214, 221, 225, 231]. The rate of formation, disintegration and accumulation of these adducts vary per nucleotide. Nucleotides

with reactive endocyclic NH (3-hmdTMP), the reactions are instantaneous, thus considered in dynamic equilibrium [214, 221, 231]. Conversely, nucleotides with exocyclic NH_2, have a slower formation rate, but hm adducts are more stable and reach higher levels of saturation, sometimes leading to an irreversible reaction, e.g. 2-hmG and 6-hmdA [213, 221, 225]. The rates of reaction are influenced by temperature, pH and ionic strength. In all cases, increase in temperature catalyses both reactions, while pH and ionic strength only significantly affect reaction rates of endocyclic adducts (with a positive correlation) [214, 231] and only marginally affect reaction rates in exocyclic adducts [215, 221]. In addition to mono-hydroxymethyl adducts, exocyclic NH_2 can form di-hydroxymethyl adducts (dhmN), such as 6-dhmdA, 4-dhmC and 2-dhmG, whoever, dhmN occur at a low rate (2-3 fold slower) and reach lower concentration (35-40 X lower) than mono adducts, but their reverse reactions is up to 100 X slower, thus more stable. For example, Shishodia *et al*. observed that while 4-dhmC degrades at 0.003 µM/s, 4-hmC degrades at 0.133 µM/s, similarly, 2-hmG degrades at 0.003 µM/s while 2-dhmG at 0.033 µM/s) [214, 215, 221].

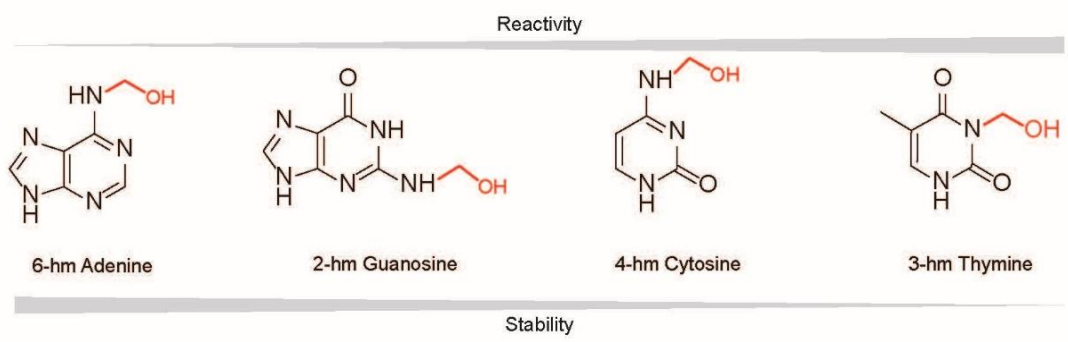


Figure 2. Sites of hydroxymethyl adduct formation in nucleobases.

In red: hydroxymethyl adducts formed by formaldehyde. 3-hmT is the most reactive base, but unstable. 6-hmA is the least reactive base but forms the most stable interactions.

Formaldehyde interaction with DNA

HCOH has been found to reversibly denature native DNA *in vitro*, destabilising both CG- and AT-rich regions equally. However, the kinetics of adduct formation are highly inhibited by base stacking, since interaction sites are involved in hydrogen bond base pairing. Therefore, for HCOH to interact with native DNA, bases need first to unstack [216]. Once the hydrogen bonds are broken and bases are unstacked, the rate

of mono-adduct addition reaction is the same as observed for mononucleotides. *In vitro* models (naked DNA) have shown the HCOH-DNA interaction initiates at thermodynamically unstable AT-rich regions with the formation of a 6-hmdA forming a weaker dA-dT bond. This bond significantly reduces the helix stability and melting temperature (Tm), inducing the denaturation of neighbouring nucleotides and enabling their reaction with HCOH. This results in the formation of methylene bridges and nucleotide crosslinks [211, 215, 218]. In these studies, crosslinked nucleotides have been found in 2 % of HCOH-fixed DNA, as inter- or intra- strand crosslinks, with a symmetric (6-hmdA-dA, 2-hmdG-dG) or asymmetric conformation (2-hmdG-dA, 6-hmdA-dC and 6-hmdG-dC) (Figure 3). Their occurrence relies on the spacing and orientation provided by the helix structure, since they form upon contact of reactive groups. Here, the most common crosslinked nucleotides found were inter-strand dA-dA crosslink, at AT-rich regions. The second most common type of DNA crosslink found is dG-dG, occur in nucleotide sequences CG, GNC and GC [217, 218, 232].

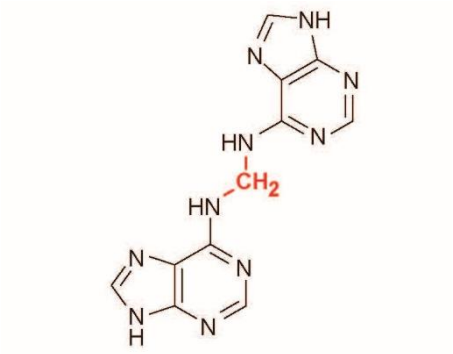


Figure 3. Example of a DNA crosslink: Adenine - Adenine dinucleotide crosslink. *Methylene bridge shown in red.*

Formaldehyde-driven DNA-Protein Crosslinks (DPC)

HCOH has been found to react highly with the side chains of Lysine (Lys), Cysteine (Cys), Histidine (His), Arginine (Arg) and Tryptophan (Trp) [219]. DPCs have been thoroughly investigated *in vitro*. Here, Lys-dG and Cys-dG were found in high yields, followed by Cys-dA, Cys-dC, and by His-dA and Trp-dG after prolonged incubations. Lys-dG is the most prominent DPC, dG reacts with Lys through the exocyclic N2 and endocyclic N1 and N3, with three possible products: *(i)* single crosslink at N2, *(ii)* a

double crosslink at N1 and N2, forming a triazinane ring, and *(iii)* a tricyclic nucleotide with a hm group at N3 (Figure 4). All species were detected even at low HCOH concentrations, but the prevalence of tricyclic structures was found to be concentration and fixation-time dependent. These were also shown to be reversible in solution, with a higher lability in single-bond structures. The second most common crosslink found was Cys-dG, forming a single and stable bond between N2 in dG and the sulfhydryl group (SH) of Cys. To a lower extent, Cys (SH) was shown to react with the N6 of dA and the N3 in dC [230, 233-235].

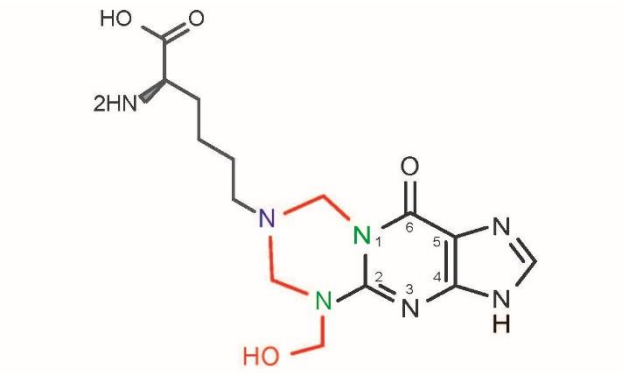


Figure 4. Triazinane ring formed by interaction of Lysine with Guanine.
Red: the methylene bridges forming the structure. Green: the reactive amino groups in Guanine. Blue: the reactive amino group in Lysine.

The cellular milieu

Despite the multiple products that can be derived from HCOH interaction *in vitro*, the conditions allowing their formation are not met *in vivo*. In the cellular milieu, the number of structures that can be generated is limited because: *(i)* lower HCOH concentrations reduce its reach; *(ii)* reactive groups in amino acids are better nucleophiles than those in nucleotides; *(iii)* HCOH does not alter tertiary structures of proteins. Thus, residues on protein surfaces or in interaction centres of native proteins provide the most accessible substrate. *(iv)* Temporal distribution is a determinant for HCOH interactions in the multi-macromolecular cellular milieu. Here, exposed groups of proximal macromolecules are more likely to interact, and interacting molecules more likely to be crosslinked together. [229, 234-237]. In line with this, the addition of Lys containing DNA binding proteins has been shown to exponentially accelerate the kinetics of the HCOH-DNA interaction, while non-DNA interacting proteins do

not crosslink with DNA nor exert any effect on this interaction. Accordingly, Lys residues are ubiquitous in DNA binding proteins, serving as mediators of DNA-Protein interactions and constitute the main crosslinked residues found in histone complexes [212, 238, 239]. These observations suggest that the HCOH-DNA interaction is facilitated or triggered by interacting DNA binding proteins. In this context, crosslinks are formed between Lys ε-NH_3 and proximal nucleotides, and after this crosslink is broken, an hm adduct persists in the nucleosides [225, 229, 238, 239]. Congruently, the most ubiquitous hmN found *in vitro* and *in vivo* is 2-hmdG, which can interact with Lys through three reactive groups [225, 230]. Surprisingly, 2-hmdG has been found to predominate in DPCs formed by other chemical agents [240] and has also been found to crosslink with aldehydes of abasic sites at opposite strands [241].

In summary, in double stranded DNA, HCOH-DNA interactions are severely limited by base stacking. Within the cellular milieu, denaturation is more likely driven by DPC formation (dG-Lys) and to lower extent by 6-dhmA formation in AT rich regions [225, 230]. This is supported by *in vivo* and *in vitro* evidence reporting favourable reaction kinetics and ubiquitous release of 2-hmdG upon crosslink reversal. Once the double helix is destabilised and nucleotides exposed, the kinetics of HCOH-DNA interaction follow that of free nucleotides, generating the same ratio of products, but to a lower extend than *in vitro*, more likely involving exocyclic NH_2 of purines, with a high prevalence of dG adducts and/or crosslinks (Table 1).

Table 1. DNA crosslinks and their frequency of occurrence *in vitro* or *in vivo* [213, 214, 217, 218, 221, 225, 232, 239, 242]

Nucleotide	Crosslinks	Type of Crosslink	Test	Prevalence	Lability
dG	dG – Lys or Cys	DNA-Protein	*in vivo*	High	High
dA	dA – Cys or His	DNA-Protein	*in vivo*	Low	High
dG	dG – Lys or Cys	DNA-Protein	*in vitro*	High	High
dA	dA – Cys	DNA-Protein	*in vitro*	Low	High
dC	dC – Cys	DNA-Protein	*in vitro*	Low	High
dG	dG – dA or dG	DNA-DNA	*in vitro*	High	Low
dA	dA – dA or dG	DNA-DNA	*in vitro*	High	Low
dG	2-hm-dG	hm adduct	*in vitro*	High	Low

dA	6-hm-dG	hm adduct	*in vitro*	High	Low
dC	4-hm-dG	hm adduct	*in vitro*	High	High

Effect of tissue processing and storage on FFPE samples

Data from the few studies available on this topic, indicate that post-fixation tissue processing is detrimental to DNA. It was found that anhydrous conditions (after dehydration) prompts molecular dehydration, which increases the formation of Schiff bases and crosslinks. Additionally, under these conditions, a fraction of hm adducts are converted into ethoxymethyl (ehm) adducts, which are more stable (10X half-time of hm adducts) and structurally bulkier, which may exacerbate depurination [243, 244]. Similarly, prolonged exposures to warm hydrocarbon solvents during paraffin embedding were found detrimental to nucleic acids [245]. In addition, storage time of FFPE specimens was shown to remarkably increase DNA damage. It has been calculated that nucleic acids from FFPE samples stored at room temperature (22 ^{o}C) reach lowest integrity values upon storage for 6 -12 months [223]. Related DNA degradation was attributed mainly to oxidation and hydrolysis, caused by residual water molecules in the sample or the environment. This is exacerbated in exposed tissue sections [246]. Accordingly, it has been shown that only changing storage conditions (low temperature and humidity) can significantly prevent degradation [223].

DNA Damage found in FFPE specimens

DNA damage as products of the above described HCOH interactions in FFPE samples have been found to be in the form of: *(i)* Crosslinks (DNA-DNA, Protein-DNA), *(ii)* depurination leading to *(iii)* DNA fragmentation and *(iv)* sequence alterations (chimeras, SNPs) [182, 228]. These accumulate with time of storage and also correlate with suboptimal fixing conditions (low pH and higher incubation times) [228, 247].

Crosslinks:

The product of the interaction of HCOH and biomolecules is the formation of crosslinks. These are ubiquitous in FFPE sample, and occur as DNA-DNA and

Protein-DNA crosslinks. As described before, these are more frequently found between dG and amino acids Lys and Cys in the form of DPCs [230, 233]. DPCs inhibit DNA amplification by blocking the processivity of DNA polymerases, terminating primer extension [248]. Despite their high prevalence in FFPE samples, it has been demonstrated that HCOH crosslinks are reversible, as seen in Figure 1, Schiff bases intermediaries can be reversed by hydration and methylene bridges formed by HCOH are heat liable with a half-life reduced from 179 h to 11.3 h upon heating from 4 ^{o}C to 47 ^{o}C [242]. In addition, the crosslink reversal reaction has been found to be influenced by pH, salt concentration and the incorporation of quenchers such as Tris-HCl. As seen in Figure 5, quenchers can sequester released HCOH in the solution, preventing the formation of additional crosslinks or act as transamination catalysts [219, 229, 249].

Heat treatment for crosslink reversal or decrosslinking is essential for DNA purification of FFPE samples and all protocols and kits for FFPE DNA purification incorporate it, typically as a 1h incubation at 90^{o}C [250]. However, recent studies have found that this high temperature incubation can lead to a high frequency of ss-breaks, and sequence artefacts, such as chimeras. It was also observed in these studies that reducing the decrosslinking temperature led to a reduction sequence artefacts, but also reduced the amount of sequencing reads (DNA available) [220, 222]. This suggests that there is still room to optimise a decrosslinking reaction conditions in order to reduce the incubation temperature, without affecting the yields of decrosslinked DNA. As mentioned before, the most abundant DNA crosslinks found, are in the form of DPCs, more frequently between DNA and DNA binding proteins[238]. These crosslinks could strain the double helix structure and promote depurination or ss-breaks. Thus, targeting these crosslinks could reduce the prevalence of sequencing artefacts.

All FFPE DNA purification methods include a protein lysis step, before decrosslinking, and decrosslinking is performed in the protein lysis buffer [251]. The efficiency of the protein would influence the dissolution of Protein-DNA crosslinks. Sodium Dodecyl Sulphate (SDS) is the protein denaturing agent of choice. SDS activity is favoured by boiling temperatures [252], which might explain 90 ^{o}C decrosslinking incubations. In addition, its ionic nature limits its interaction with DNA and does not denature DNA. The efficiency of lysis buffers on FFPE tissues and their

impact in decrosslinking has yet to be investigated. It has been suggested that utilising chaotropes might improve the quality of yielded nucleic acids [254].

Chaotropes, such as Guanidium hydrochloride (GuHCL) are among the most potent protein denaturants [255]. The high denaturing activity of GuHCL is due to its ability to associate with different protein groups, including the carboxylic groups, non-polar hydrophobic groups (through hydrophobic interactions), and polar side chains (negatively or positively charged Arginine). This activity is also due to their capacity to self-associate, despite electrostatic repulsion and their high affinity for water molecules [256]. This reduces the likelihood for proteins to self-associate and reduces the penalty for unfolding in the presence of water [257]. Furthermore, unlike SDS, chaotropes interacts with nucleic acids, altering their secondary and tertiary structure [258, 259]. In fact, 1M concentrations of GuHCL have been shown to reduce the melting temperature of DNA by 13^{o}C, and increase the stringency of its hybridisation, promoting correct base pairing[260]. Similar concentrations of GuHCL have been shown to increase the activity of Proteinase K and other Proteinases [261, 262], which might be facilitated by the ability of GuHCL to increase the torsional mobility of denatured proteins [263], thus facilitating access to Proteinase while also protecting the DNA structure at DPC sites. In addition, GuHCL activity is not affected by temperature [264]. Altogether, these features might enable a lower decrosslinking temperature. Furthermore, while interactions between the Guanylyl groups in GuHCL and HCOH have not been studied in these settings, these have been reported under different experimental conditions [265]

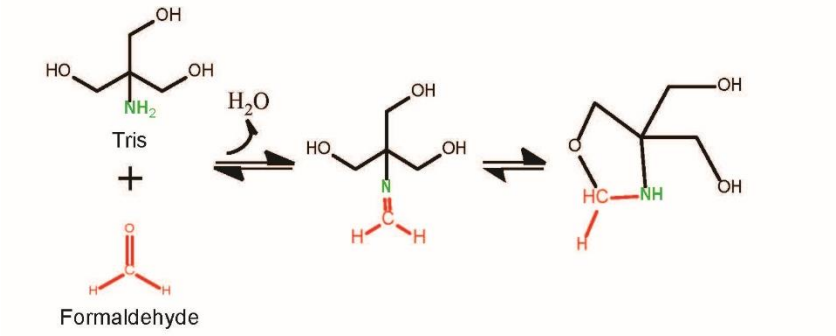


Figure 5. Tris as a scavenger of HCOH.

Reactive amino groups in Tris (green), interact with HCOH (red) sequestering it from the solution. Tris is used as a HCOH quencher.

Depurination:

N-glycosylic bonds in DNA are labile to spontaneous hydrolysis rendering (Apurinic/Apyrimidinic) AP sites. This has been observed to occur at a rate 10 times higher in purines than pyrimidines. In fact, depurination is the main route for DNA degradation in DNA deprived of DNA repair mechanisms [266]. Depurination is heat activated and highly susceptible to low pH and salt concentration. Furthermore, depurination is accelerated by modification of purines bases comparable with those induced by HCOH [210, 266]. Similarly, anhydrous DNA, similar to DNA in FFPE samples, has been shown to undergo spontaneous depurination. This is due to residual endogenous (intramolecular) and exogenous (environmental humidity) water, pH memory and a possible catalysis driven by orthophosphates (from buffered formalin). In addition, upon dehydration, the secondary structure of DNA is distorted to a conformation (from B DNA to A DNA) more prone to denaturation, which, coupled with bulky structures found in HCOH fixed DNA, facilitates the breakage of labile N-glycosylic bonds [243, 267, 268]. Abasic sites dramatically weaken the double-helix, inducing its denaturation and exposing nucleotides to HCOH. Additionally, open-chain aldehyde residues in abasic site have been found to crosslink with dG and dA or undergo spontaneous β-elimination forming ss breaks (Figure 6) [209, 224, 241, 269-271]. Finally, abasic sites can obscure DNA analysis. DNA polymerases have very low tolerance to abasic sites and stall replication upon their encounter, thus inhibiting PCR. In the rare event of bypassing them, they would either favour the incorporation of dA (A-rule) or produce frame-shifts [272-276]. Given the ubiquity of these lesions in metabolically active cells (10,000 events/day) these are efficiently repaired by ubiquitous AP (Apurinic) endonucleases, which initiate the base excision repair (BER) pathway [266]. The *in vitro* recreation of this pathway might offer a solution for repair of these lesions in FFPE samples.

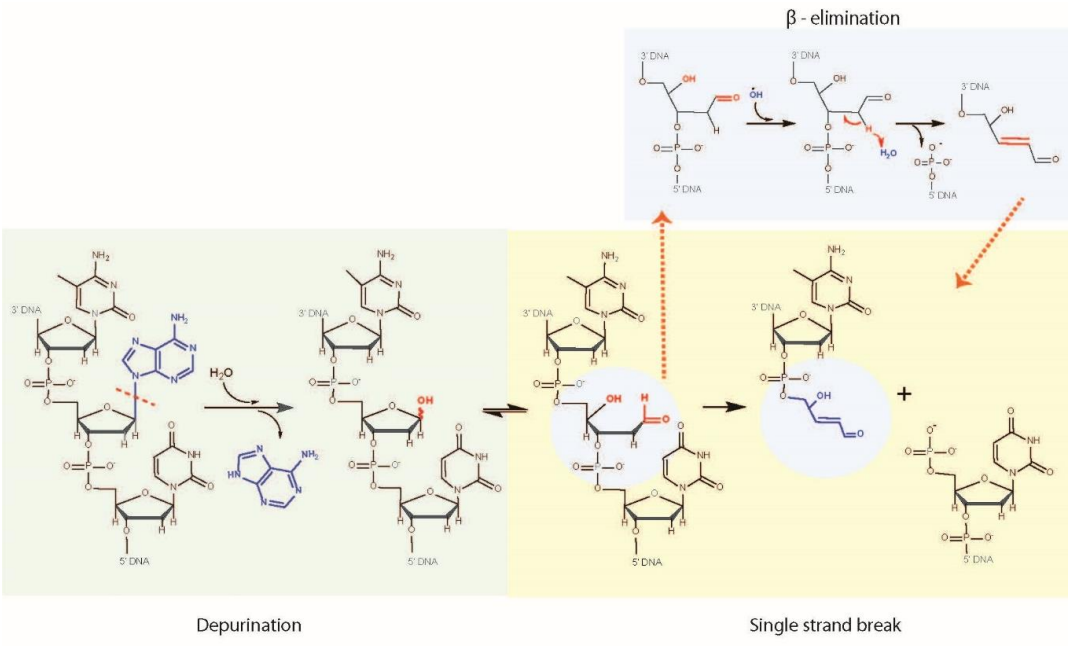


Figure 6. Depurination leading to ss-breaks.

Chemical interactions leading to spontaneous strand breaks in depurinated sites.

Fragmentation:

DNA fragmentation has been widely reported in FFPE samples, and fragments above 300 bp are unlikely to amplify under standard PCR reaction conditions [173, 277]. In fact, protocols for DNA analysis of FFPE specimens are usually developed to target fragments below or equal to 200 bp [172]. DNA fragmentation is age-dependent, accumulating over time and is accelerated by poor fixation practices [244, 247]. As described above, abasic sites lead to the formation of single-strand (ss) breaks that if located within 10 bp of an opposite ss-break will turn into double strand (ds) break [266, 269]. It has been estimated that 18 ss breaks per day occur in archival samples. At this rate, models predict an average fragment length drop from 1 Mbp to < 2 Kbp in 5 years [247, 278]. This fits the time-depended fragmentation of FFPE samples, in which samples fixed for over 5 years appear as smears of less than 1.5 Kbp [223, 277]. While there is no repair mechanism that could faithfully correct ds breaks for unknown genomic targets *in vitro*, the repair of ss breaks has been achieved via the BER system and DNA mismatch repair (MMR) system. In addition, the reduction of ss breaks in FFPE has also been investigated by regulating the temperature and time of decrosslinking incubation [279, 280].

Sequence artefacts:

The quality of sequencing analysis is affected by sequence artefacts found in FFPE samples, as evidenced by the reduction in sequencing depth and uniformity, shorter fragments read, reduced ratio of pass-filter reads, high number of chimeric reads and reduced GC ratios. FFPE derived artefacts are also present in the form of single nucleotide polymorphisms (SNPs), translocations, and insertions and deletions (indels) [222].

Deamination of Cytosine. C:G > T:A transitions are a product of deamination of dC to dU or 5-methyl deoxycytosine (5m-dC) to dT (Figure 6) in CpG regions. Early genomic studies on FFPE tumour specimens found disproportionally high rates of C:G > T:A transitions that were thought to be FFPE derived artefacts [228, 281]. As a consequence, use of Uracil-DNA glycosylase (UDG) to remove Uracil moieties was proposed, to improve the sequencing quality of FFPE DNA [282]. However, improvements were questionable, with only marginal reductions of C > T SNPs [283-285], which was partially attributed to higher rates of C > T originating from 5m-dC in CpG sites [284, 286, 287].

The occurrence of these artefacts was recently clarified by whole genome/exome sequencing studies using paired FFPE/Frozen tissues of non-cancerous and cancerous origin. These studies revealed that C > T transitions occur at much lower rates than previously proposed, and demonstrated that a large extent of SNPs disparities were products of intra-tumour or sampling heterogeneity and not FFPE derived [220, 222, 250, 288]. Remarkably, these studies confirmed that artefacts derived from oxidative stress (i.e. G > A) DNA damage are unique to FFPE samples [289]. Further studies using similar experimental conditions revealed that sample age, ischemic time and fixation conditions were also confounding for these disparities, and low template DNA input and shallow sequencing analysis were shown to be exacerbating factors [290, 291].

Congruently, dU is a common DNA base damage in metabolically active cells [292]. In this context, dU ubiquity is largely attributed to DNA metabolism and misincorporation events during replication. However, dU also arises from the deamination of an inherently unstable dC. This is reflected in the multifaceted dU

repair system comprising 4 DNA glycosylases in mammals (2 in bacteria) [276, 293, 294]. Of particular interest for FFPE samples, which exhibit oxidation patterns, is the oxidation of dC bases. It has been shown that these lesions yield unstable species that are easily deaminated, such as 5-hydroxydC (5-OHdC), 5-hydroxydU (5-OHdU) and uracil-glycol (Figure 7). These lesions have been found to be bypassed by polymerases leading to C > T and C > G SNPs, the most common SNPs found in FFPE samples. These damaged nucleotides are targeted by DNA glycosylases shown in Figure 7. Thus, it is worth investigating the reconstitution of BER using these DNA glycosylases that target oxidised and deaminated Cytosine (Fpg, Endo III, Endo V and Endo VIII) to reduce the rate of these SNPs and improve the overall sequencing quality of FFPE samples.

Deamination of 5-methyl cytosine. Also congruent to the above studies is the fact that methylated dC is 3–4 fold more labile to deamination than dC. In mammals, 5m-dC is present at high levels (1% of the genome) at CpG sites, where 80-90 % of dC are methylated. CpG are mutational hotspots, with a remarkably high rate of spontaneous C > T SNPs [295]. Their high rate of occurrence can be partially explained by the trade-off between their vestigial role in preventing the integration of transposable elements and their recent evolution in complex vertebrates for gene regulation [296]. Here, while dU is targeted by 4 glycosylases, dT arising from 5-mdC deamination is only targeted directly by Thymidine DNA glycosylase (TDG) (Figure 7) [270, 295], even though the recognition and excision of this lesion is far more complex than that of non-canonical dU. It has been recently revealed that in the cytosine demethylation pathway, intermediate products such as 5-hydroxymethylcytosine (5-OHmdC) are oxidised to target their removal by TDG. These oxidised intermediates are randomly distributed across CpG sites, further obscuring the discrimination of FFPE derived artefacts. Given the high rate of SNPs naturally occurring in the biological context, it is difficult to draw conclusions on the influence of FFPE in the deamination of 5-mdC. The processivity of TDG is enhanced by Neil (Endonuclease VIII in bacteria), which has overlapping substrates with TDG (oxidised pyrimidines) [297].This encourages the use of these DNA glycosylases to target oxidised and deaminated products of 5-mdC.

Oxidative Damage. Variable levels of SNPs derived from oxidative DNA damage have been identified in FFPE samples, frequently in the form of C:G > A:T, A:T > G:C, T:A > C:G [289]. As explained in previous sections, dG provides the most favourable template for HCOH interactions forming very stable products, resulting in a high crosslinking reactivity [230, 234]. This is reflected in a high rate of oxidised dG adducts in FFPE samples, where it has been found at 4 -7 X higher ratios then in paired frozen controls [298]. These observations have been confirmed by *in vivo* studies demonstrating a marked increase in biomarker reactive oxygen species known to induce dG oxidation upon HCOH exposure [299]. In agreement with this, oxidative damage is the main form of DNA damage in metabolically active cells, and more than 50 different oxidised DNA base modifications have been identified. dG has the lowest redox potential, and is therefore the most frequently oxidised base [300], and the most frequently observed dG (8-oxo-dG) is considered a marker for oxidative stress damage (Figure 7). The biological relevance of these lesions is also confirmed in bacteria by the complex and robust systems that repair them. Here, oxidative damage is targeted by DNA glycosylases capable of DNA backbone cleavage, increasing the repair processivity. These enzymes also have overlapping substrate specificities, thus ensuring substrate repair even when one is missing (Figure 7) [294, 301].*Chimeras.*

Finally, new studies have provided compelling evidence indicating that, to a large extent, structural rearrangement chimeras (indels and translocations) found in FFPE samples, are formed during library preparations with methods that include end-repair (with T4 DNA ligase) and by high temperature incubation used for decrosslinking DNA. It was found that by annulling and/or modifying this treatments, these were significantly reduced [220].

Finally, it has been shown that sequence alterations in FFPE samples do not interfere with cancer diagnostics in clinical settings as long as optimal conditions are in place: high input (> 250 ng) template DNA, high quality of DNA (scored through a quality assessment) and high sequencing depth (8X). However, when these conditions are not met, the rate of sequence artefacts can be detrimental to the analysis. Here, the number of reads and the overall sequencing coverage are significantly reduced and the frequency of SNPs can increase to 1/1000 bp and lead to erroneous sequence analysis [286, 290, 302, 303]. Under these conditions, the quality of the analysis will be more

reliant on optimised methods for DNA purification, DNA quality control, library preparation, sequencing analysis and the incorporation of DNA repair to the workflow [250, 302-304].

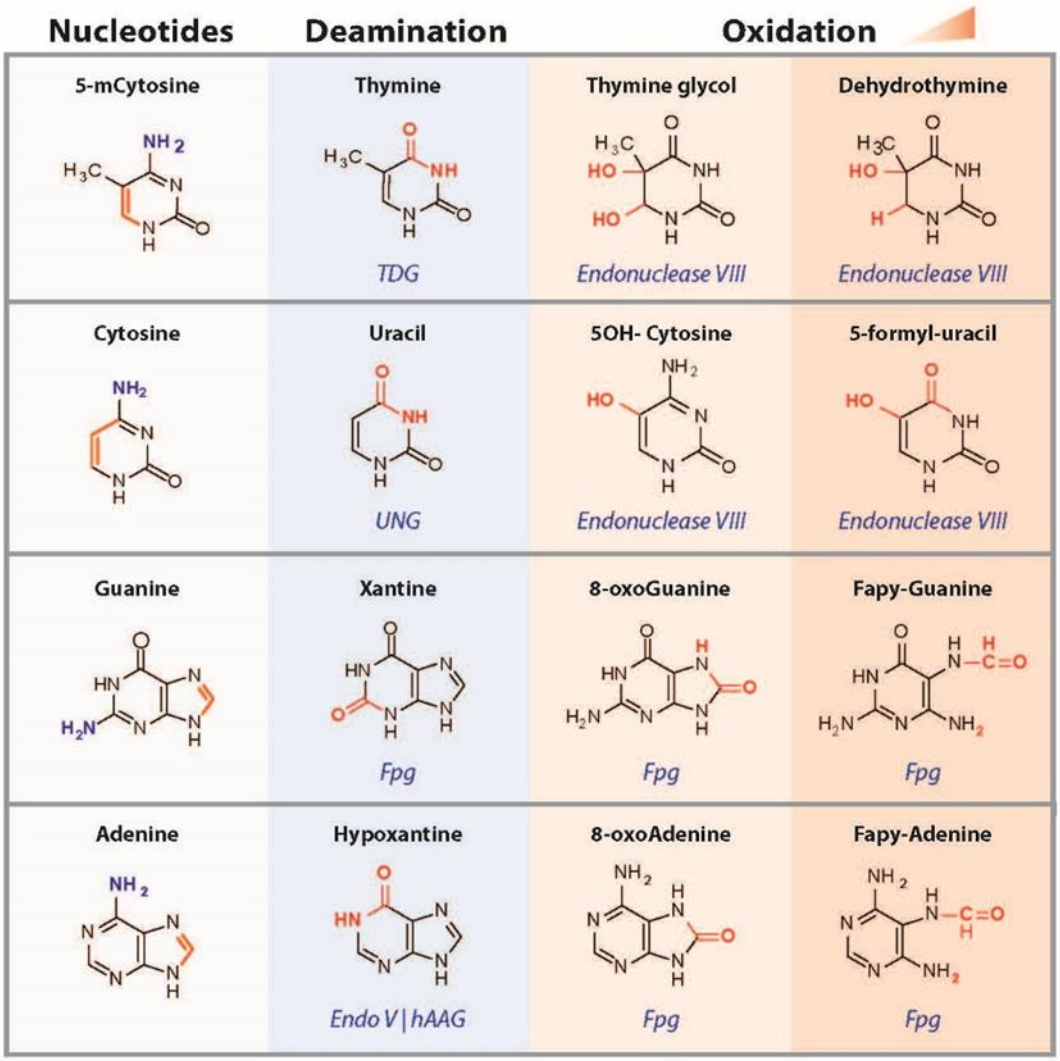


Figure 7. Common base lesions caused by formalin fixation and the BER DNA glycosylases that target them.

FFPE Effects on Bacterial DNA

Despite the plethora of accrued knowledge on FFPE in mammalian/human DNA, very little is known on the effects of HCOH fixation on bacterial DNA [305]. In principle, bacterial DNA will interact with HCOH as described in *in vitro* assays. However, there are certain differences in the conformation and packaging of bacterial DNA, as well

as differences in methylation pattern and replication and transcription rates (less single stranded DNA) that might influence the rate of this interaction. For example, the small sized bacterial chromosome is in most cases circular and packaged in set of independent supercoiled domains with uncoupled topological states. Hence, a depurination event in one domain will have very little effect on another [306, 307]. This, coupled with the low rate of depurination events observed for bacteria (in *E. coli* – one occurring every 2 generations), reduces the likelihood of depurination driven DNA damage [210]. Conversely, the supercoiled DNA structure is held by an array of histone-like proteins and has been found to interact with proteins localised all over the cell, including the cellular envelope [306], likely to facilitate HCOH-DNA interactions and crosslink formation.

Likewise, while C > T transitions might be the most common reported SNPs in human cancerous DNA, this might differ in bacteria. As explained above, a higher proportion of C > T SNPs in humans derive from deamination of 5m-dC. However, in bacteria 5m-dC are constrained to unique very short-patch repair (VSR) sequences at a log fold lower rate than observed in eukaryotes and signalling methylation is mostly done through amino groups of dA [308]. Congruently, spontaneous deamination in eukaryotic cells occurs at a 40-fold higher rate than in bacteria [266]. Altogether, the differences between structure and dynamics of bacterial and eukaryotic genomes raise many queries on the HCOH interaction with the bacterial genome. A better understanding of this interaction might lead to a clearer path to pursue microbiome analysis of these specimens. Despite the lack of this foundational knowledge, the targeted detection of microbial DNA in FFPE tissues has already proven feasible for the detection of pathogens by PCR and, to a certain degree, by 16S sequencing [199, 202, 309]. In addition, they are now serving as templates for microbial barcoding surveys [199, 204, 205].

Considerations for sequence-based analysis of bacterial FFPE samples:

As with investigations into FFPE induced DNA damage or repair strategies, the glut of research carrying out sequence-based analysis of FFPE human should be used as a resource to attempt to mitigate errors in the analysis of bacterial DNA. If the effects

of FFPE on bacterial DNA are similar to those in human DNA, then at a minimum the following can be expected. 1) *Fragmentation:* The fragmentation of DNA will impact sequencing strategies. Long read single molecule real time sequencing strategies may not be a viable option in these circumstances, and depending on the severity of the fragmentation, the popular amplicon sequencing strategies employed to characterise bacterial communities may be affected. As an example, the variable regions most commonly targeted in 16S rRNA gene sequencing (V3-V4) have a combined length of ~460 bp on average, which may exceed the fragment length of much of the DNA in an FFPE sample, reducing the already limited amount of DNA available. DNA fragment length should be assessed in advance and choices on amplicons and/or sequencing chemistry dictated by this.

DNA damage: In the field of human cancer genomics, there is the recurring problem of differentiating low frequency genetic variants present in a small number of cancerous cells from decoys resulting from DNA damage[226]. The anticipated effect transposed onto bacterial genomic research is an increase in mutations possibly leading to erroneous speciation events. To combat this, there is further potential to take advantage of existing research if it can be successfully extrapolated into bacterial research. Bioinformatics tools have been developed to differentiate between true low frequency variants and artefacts of the FFPE process [310], and with sufficient FFPE bacterial DNA sequence data available, approaches such as this could be adapted.

DNA Repair

The Base Excision Repair (BER) system is the main cellular pathway for repair of lesions that do not cause significant distortions in the DNA helical structure, such as damaged bases, AP sites and ss breaks. This pathway is well conserved across evolution. The enzymatic repair of DNA by the BER pathway consists of five basic steps: (*i)* base excision by a DNA glycosylase (*ii)* Backbone incision by AP lyase, (*iii)* Ends processing by a polynucleotide kinase or 3’ – 5’ exonuclease, (*iv)* Gap filling by a polymerase, and *(v)* Nick ligation by a ligase [294, 300].

BER is initiated by the recognition and removal of damaged base by a DNA glycosylases [311]. These identify and remove damaged nucleotides by base-flipping

(Figure 8), which allows the insertion of the base onto a recognition pocket that holds the active cleavage site. The affinity at which recognition pockets bind their target dictates the target selectivity; for example, glycosylases recognising dU bind tighter to their targets, while those recognising oxidative damage are looser and able to process a wider range of damage [312, 313]. Similarly, those targeting uracil lesions only operate as glycosylases (monofunctional), while those targeting oxidative damage also operate as AP lyase (bifunctional) and the downstream BER pathways will be dictated by these modes of action (Figure 8) [294, 300, 313].

In bacteria, for monofunctional DNA glycosylases, the AP site yielded after base removal is excised by an AP endonuclease (Endo IV/ Exo III), by cleaving the phosphodiester bond 5' of the AP site, generating a nick with deoxyribose phosphate (dRP) residue in the 5' terminus and a clean 3'OH terminus (Figure 9A). The 5'dRP residue can be removed by the 3' – 5' exonuclease activity of DNA polymerase during strand displacement, where 2-8 nucleotides downstream are displaced, thus leading to long-patch BER sub-pathway. The filled gap is later sealed by DNA Ligase I (Figure 8) [292, 294, 300, 311-313]. For bifunctional glycosylases, after base excision, the ribose cleavage by β-elimination yields a Phospho-α,β-unsaturated aldehyde residue at the 3' end (Figure 9B). This residue can be either removed by an AP endonuclease, leading to long-patch repair, or in the case of β/δ – glycosylases, this residue will be removed by the cleavage of the phosphate-ribose bond by the glycosylase, and yields a 3' end phosphate. This blocking phosphate is later removed by either an AP endonuclease or a polynucleotide kinase (PNK) and the lesion filled through short-patch BER and sealed with Ligase [294, 300, 301, 311-314] (Figure 9B.)

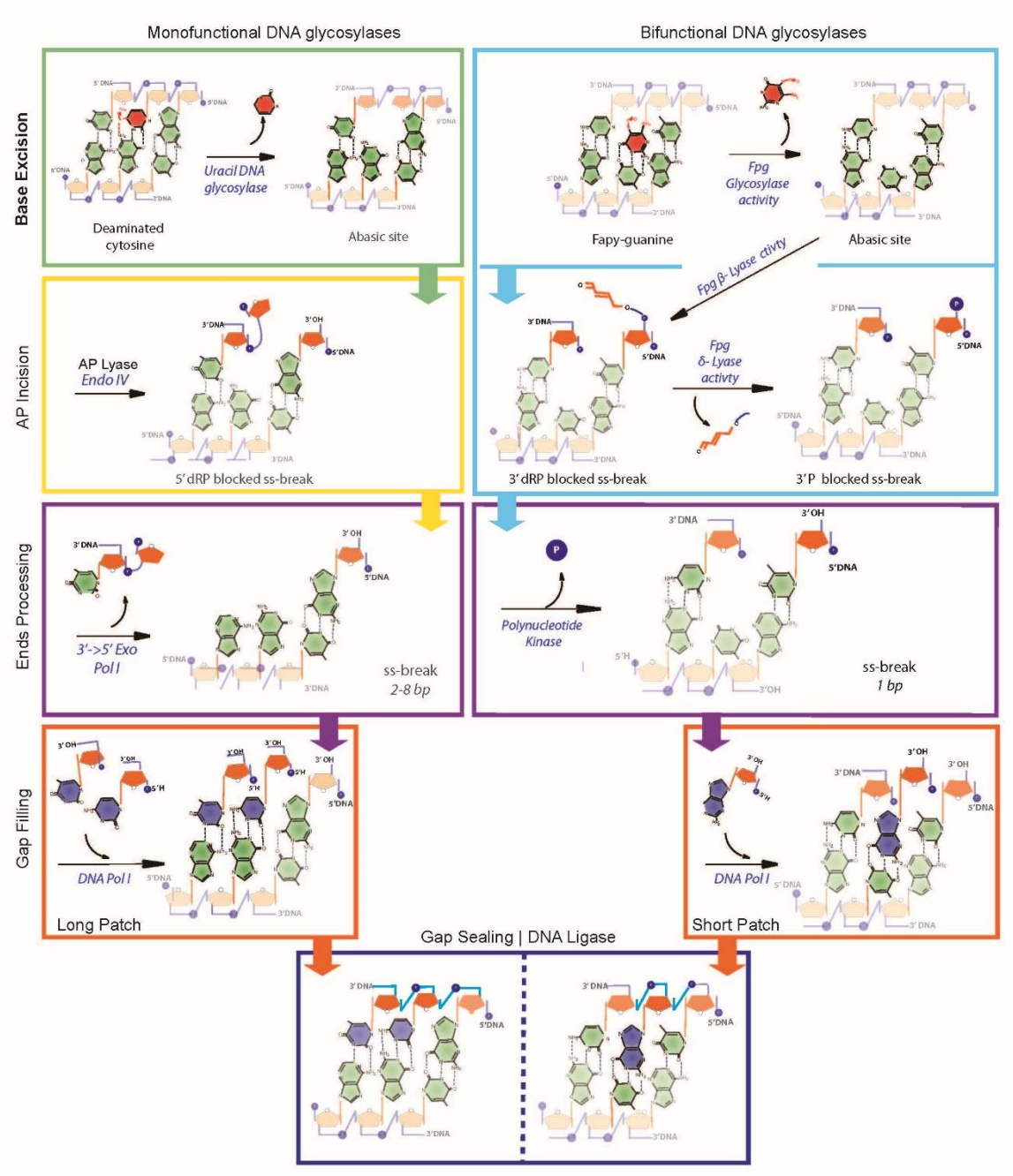


Figure 8. BER sub pathways.

The 5 steps for BER reaction illustrated: (1) base excision by a DNA glycosylase; (2) AP incision by an AP lyase/AP endonuclease; (3) Ends processing by polynucleotide kinase or Polymerase. (4) Gap filling by a polymerase. (5) Gap sealing by Ligase

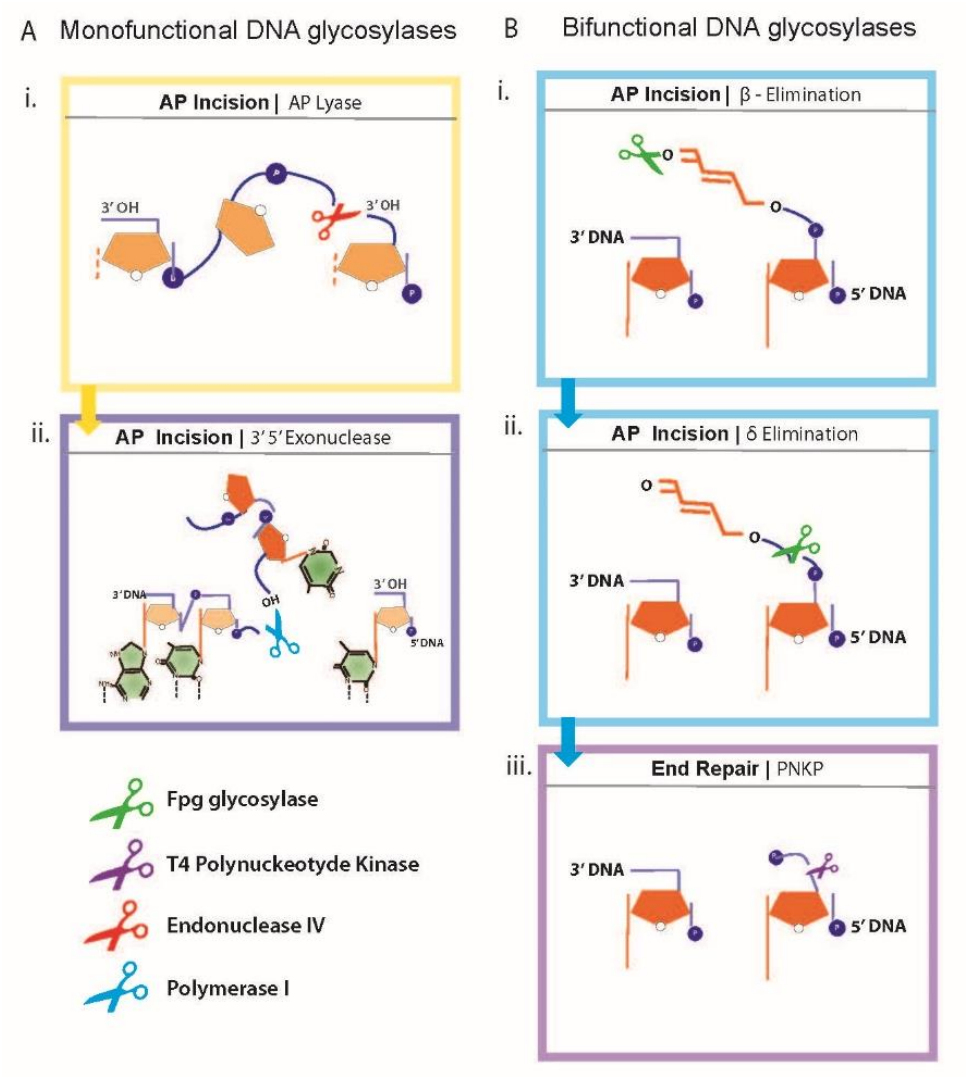


Figure 9. Base excision and end repair.

End repair driven by mono- or bi- functional DNA glycosylases.

Concluding Remarks

Only through a holistic understanding of FFPE-induced DNA damage can methods to repair relevant damage, thus improving the quality of genomic and microbial sequencing analyses, be devised. The latest developments in DNA repair and reconstitution of the BER pathway offer a unique opportunity to achieve this. Overall, a deepening in appreciation of the impacts that FFPE-induced DNA damage has on various analyses of DNA stands to benefit genomics and microbiome research alike.

REFERENCES

1. Clarke, L.J. and R.I. Kitney, *Synthetic biology in the UK – An outline of plans and progress.* Synthetic and Systems Biotechnology, 2016. **1**(4): p. 243-257.

2. Oldham, P., S. Hall, and G. Burton, *Synthetic biology: mapping the scientific landscape.* PLoS One, 2012. **7**(4): p. e34368.

3. Si, T. and H. Zhao, *A brief overview of synthetic biology research programs and roadmap studies in the United States.* Synthetic and Systems Biotechnology, 2016. **1**(4): p. 258-264.

4. OECD, *The bioeconomy to 2030 : designing a policy agenda* 2009, Paris: Organization for Economic Co-operation and Development (OECD) [Paris].

5. Joyce, S., A.-M. Mazza, and S. Kendall, *Positioning synthetic biology to meet the challenges of the 21st Century* in *summary report of a six academies symposium series*, T. National Academy of Engineering; National Research Council (US); Policy and Global Affairs; Division on Earth and Life Studies; Committee on Science, and Law; Board on Life Sciences, Editor. 2013, The National Academies Press,: Washington, D.C. p. 80p.

6. Boeke, J.D., et al., *GENOME ENGINEERING. The Genome Project-Write.* Science, 2016. **353**(6295): p. 126-7.

7. Kosuri, S. and G.M. Church, *Large-scale de novo DNA synthesis: technologies and applications.* Nat Methods, 2014. **11**(5): p. 499-507.

8. Endy, A.E.D., *Synthetic Biology: What it is and why it matters?*, in *Synthetic aesthetics : investigating synthetic biology's designs on nature*, A.D. Ginsberg, Editor. 2014, MIT Press: Cambridge, Mass. p. xxii, 349 p.

9. Ledford, H., *CRISPR, the disruptor.* Nature, 2015. **522**(7554): p. 20-4.

10. Carlson, R., *Estimating the biotech sector's contribution to the US economy.* Nat Biotechnol, 2016. **34**(3): p. 247-55.

11. Booth, B.L., *This time may be different.* Nature Biotechnology, 2016. **34**(1): p. 25-30.

12. Limited, E.Y.G., *Biotechnology Report 2016: Beyond borders - Returning to Earth*. 2016: London, UK. p. 83.

13. Bergin, J., *Synthetic Biology: Global Markets*. 2017, BCC Research: Synthetic Biology: Global Markets: Wellesley, MA, USA. p. 270.

14. Manyika, J., et al., *Disruptive technologies: Advances that will transform life, business, and the global economy*. 2013, McKinsey Global Institute. p. 176.

15. Baldwin, G., *Synthetic biology : a primer*. 2012, Singapore ; London: Imperial College Press ; distributed by World Scientific. xiv, 179 p.

16. Cederbaum, S.D., et al., *Recombinant DNA in medicine.* West J Med, 1984. **141**(2): p. 210-22.

17. Keen, H., et al., *Human insulin produced by recombinant DNA technology: safety and hypoglycaemic potency in healthy men.* Lancet, 1980. **2**(8191): p. 398-401.

18. Paddon, C.J. and J.D. Keasling, *Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development.* Nat Rev Microbiol, 2014. **12**(5): p. 355-67.

19. Keasling, J.D., *Synthetic biology and the development of tools for metabolic engineering.* Metab Eng, 2012. **14**(3): p. 189-95.

20. Lee, Y.J. and K.J. Jeong, *Challenges to production of antibodies in bacteria and yeast.* J Biosci Bioeng, 2015. **120**(5): p. 483-90.

21. Zarco, M.F., T.J. Vess, and G.S. Ginsburg, *The oral microbiome in health and disease and the potential impact on personalized dental medicine.* Oral Dis, 2012. **18**(2): p. 109-20.

22. Chen, H. and W. Jiang, *Application of high-throughput sequencing in understanding human oral microbiome related with health and disease.* Front Microbiol, 2014. **5**: p. 508.

23. Al-Ghazzewi, F.H. and R.F. Tester, *Biotherapeutic agents and vaginal health.* J Appl Microbiol, 2016. **121**(1): p. 18-27.

24. Kong, H.H., et al., *Performing Skin Microbiome Research: A Method to the Madness.* J Invest Dermatol, 2017. **137**(3): p. 561-568.

25. Schwabe, R.F. and C. Jobin, *The microbiome and cancer.* Nat Rev Cancer, 2013. **13**(11): p. 800-12.

26. Viaud, S., et al., *Gut microbiome and anticancer immune response:*. Cell Death Differ, 2014.

27. Sivan, A., et al., *Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy.* Science, 2015. **350**(6264): p. 1084-9.

28. Limaye, S.A., et al., *Phase 1b, multicenter, single blinded, placebo-controlled, sequential dose escalation study to assess the safety and tolerability of topically applied AG013 in subjects with locally advanced head and neck cancer receiving induction chemotherapy.* Cancer, 2013. **119**(24): p. 4268-76.

29. Byrne, W.L. and M. Tangney, *Bacteria as Gene Therapy Vectors for Cancer*, in *Gene and Cell Therapy: Therapeutic Mechanisms and Strategies*, N. Smyth Templeton, Editor. 2015, CRC Press.

30. Schmidt, I., R.J.M. van Spanning, and M.S.M. Jetten, *Denitrification and ammonia oxidation by Nitrosomonas europaea wild-type, and NirK- and NorB-deficient mutants.* Microbiology, 2004. **150**(12): p. 4107-4114.

31. Murphy, C., et al., *Intratumoural production of TNFα by bacteria mediates cancer therapy.* PLOS ONE, 2017. **12**(6): p. e0180034.

32. Vandenbroucke, K., et al., *Orally administered L. lactis secreting an anti-TNF Nanobody demonstrate efficacy in chronic colitis.* Mucosal Immunology, 2009. **3**: p. 49.

33. Deplazes, A., *Piecing together a puzzle. An exposition of synthetic biology.* EMBO Rep, 2009. **10**(5): p. 428-32.

34. Endy, D., *2003 Synthetic Biology Study*, in *2003 Synthetic Biology Study*. 2007, MIT. p. 18.

35. Endy, D., *Foundations for engineering biology.* Nature, 2005. **438**(7067): p. 449-53.

36. Elowitz, M.B. and S. Leibler, *A synthetic oscillatory network of transcriptional regulators.* Nature, 2000. **403**(6767): p. 335-8.

37. Gardner, T.S., C.R. Cantor, and J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli.* Nature, 2000. **403**(6767): p. 339-42.

38. Newcomb, J., Carlson, R., S. C. Aldrich, *Genome Synthesis and Design Futures: Implications for the U.S. Economy*. 2007, Bio-Economic Research Associates (BIO-ERA). p. 172.

39. Flores Bueso, Y. and M. Tangney, *Synthetic Biology in the Driving Seat of the Bioeconomy.* Trends Biotechnol, 2017.

40. Lee, S.K., et al., *Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels.* Current Opinion in Biotechnology, 2008. **19**(6): p. 556-563.

41. Gupta, P. and S.C. Phulara, *Metabolic engineering for isoprenoid-based biofuel production.* J Appl Microbiol, 2015. **119**(3): p. 605-19.

42. Wang, B.W., et al., *Branched-chain higher alcohols.* Adv Biochem Eng Biotechnol, 2012. **128**: p. 101-18.

43. Huffer, S., et al., *Escherichia coli for biofuel production: bridging the gap from promise to practice.* Trends Biotechnol, 2012. **30**(10): p. 538-45.

44. Vickers, C.E., L.M. Blank, and J.O. Kromer, *Grand challenge commentary: Chassis cells for industrial biochemical production.* Nat Chem Biol, 2010. **6**(12): p. 875-7.

45. Khalil, A.S. and J.J. Collins, *Synthetic biology: applications come of age.* Nat Rev Genet, 2010. **11**(5): p. 367-79.

46. Petzold, C., et al., *Analytics for metabolic engineering.* Frontiers in Bioengineering and Biotechnology, 2015. **3**(135).

47. Liu, R., et al., *Genome scale engineering techniques for metabolic engineering.* Metabolic Engineering, 2015. **32**: p. 143-154.

48. Slusarczyk, A.L., A. Lin, and R. Weiss, *Foundations for the design and implementation of synthetic genetic circuits.* Nat Rev Genet, 2012. **13**(6): p. 406-20.

49. Andrianantoandro, E., et al., *Synthetic biology: new engineering rules for an emerging discipline.* Molecular Systems Biology, 2006. **2**(1).

50. Wang, Y.H., K.Y. Wei, and C.D. Smolke, *Synthetic biology: advancing the design of diverse genetic systems.* Annu Rev Chem Biomol Eng, 2013. **4**: p. 69-102.

51. Cameron, D.E., C.J. Bashor, and J.J. Collins, *A brief history of synthetic biology.* Nat Rev Microbiol, 2014. **12**(5): p. 381-90.

52. Esvelt, K.M. and H.H. Wang, *Genome-scale engineering for systems and synthetic biology.* Mol Syst Biol, 2013. **9**: p. 641.

53. Cambray, G., V.K. Mutalik, and A.P. Arkin, *Toward rational design of bacterial genomes.* Curr Opin Microbiol, 2011. **14**(5): p. 624-30.

54. Purnick, P.E. and R. Weiss, *The second wave of synthetic biology: from modules to systems.* Nat Rev Mol Cell Biol, 2009. **10**(6): p. 410-22.

55. Mutalik, V.K., et al., *Precise and reliable gene expression via standard transcription and translation initiation elements.* Nat Methods, 2013. **10**(4): p. 354-60.

56. Whitaker, W.R., et al., *Engineering robust control of two-component system phosphotransfer using modular scaffolds.* Proc Natl Acad Sci U S A, 2012. **109**(44): p. 18090-5.

57. Keung, A.J., et al., *Chromatin regulation at the frontier of synthetic biology.* Nat Rev Genet, 2015. **16**(3): p. 159-171.

58. Brophy, J.A. and C.A. Voigt, *Principles of genetic circuit design.* Nat Methods, 2014. **11**(5): p. 508-20.

59. Dominguez, A.A., W.A. Lim, and L.S. Qi, *Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation.* Nat Rev Mol Cell Biol, 2016. **17**(1): p. 5-15.

60. Dueber, J.E., et al., *Reprogramming control of an allosteric signaling switch through modular recombination.* Science, 2003. **301**(5641): p. 1904-8.

61. Bonnet, J., et al., *Amplifying genetic logic gates.* Science, 2013. **340**(6132): p. 599-603.

62. Hasty, J., D. McMillen, and J.J. Collins, *Engineered gene circuits.* Nature, 2002. **420**(6912): p. 224-30.

63. Ozbudak, E.M., et al., *Regulation of noise in the expression of a single gene.* Nat Genet, 2002. **31**(1): p. 69-73.

64. Becskei, A. and L. Serrano, *Engineering stability in gene networks by autoregulation.* Nature, 2000. **405**(6786): p. 590-3.

65. Green, Alexander A., et al., *Toehold Switches: De-Novo-Designed Regulators of Gene Expression.* Cell, 2014. **159**(4): p. 925-939.

66. Isaacs, F.J., et al., *Engineered riboregulators enable post-transcriptional control of gene expression.* Nat Biotechnol, 2004. **22**(7): p. 841-7.

67. Bayer, T.S. and C.D. Smolke, *Programmable ligand-controlled riboregulators of eukaryotic gene expression.* Nat Biotechnol, 2005. **23**(3): p. 337-43.

68. Looger, L.L., et al., *Computational design of receptor and sensor proteins with novel functions.* Nature, 2003. **423**(6936): p. 185-90.

69. Kramer, B.P., et al., *An engineered epigenetic transgene switch in mammalian cells.* Nat Biotechnol, 2004. **22**(7): p. 867-70.

70. Guet, C.C., et al., *Combinatorial synthesis of genetic networks.* Science, 2002. **296**(5572): p. 1466-70.

71. Kaern, M., W.J. Blake, and J.J. Collins, *The engineering of gene regulatory networks.* Annu Rev Biomed Eng, 2003. **5**: p. 179-206.

72. Atkinson, M.R., et al., *Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli.* Cell, 2003. **113**(5): p. 597-607.

73. Guntas, G. and M. Ostermeier, *Creation of an allosteric enzyme by domain insertion.* J Mol Biol, 2004. **336**(1): p. 263-73.

74. Park, S.H., A. Zarrinpar, and W.A. Lim, *Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms.* Science, 2003. **299**(5609): p. 1061-4.

75. Bulter, T., et al., *Design of artificial cell-cell communication using gene and metabolic networks.* Proc Natl Acad Sci U S A, 2004. **101**(8): p. 2299-304.

76. You, L., et al., *Programmed population control by cell-cell communication and regulated killing.* Nature, 2004. **428**(6985): p. 868-71.

77. Basu, S., et al., *Spatiotemporal control of gene expression with pulse-generating networks.* Proc Natl Acad Sci U S A, 2004. **101**(17): p. 6355-60.

78. Siuti, P., J. Yazbek, and T.K. Lu, *Synthetic circuits integrating logic and memory in living cells.* Nat Biotechnol, 2013. **31**(5): p. 448-52.

79. Borkowski, O., C. Gilbert, and T. Ellis, *SYNTHETIC BIOLOGY. On the record with E. coli DNA.* Science, 2016. **353**(6298): p. 444-5.

80. Tamsir, A., J.J. Tabor, and C.A. Voigt, *Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'.* Nature, 2011. **469**(7329): p. 212-5.

81. Din, M.O., et al., *Synchronized cycles of bacterial lysis for in vivo delivery.* Nature, 2016. **536**(7614): p. 81-5.

82. Tabor, J.J., et al., *A synthetic genetic edge detection program.* Cell, 2009. **137**(7): p. 1272-81.

83. Teague, B.P. and R. Weiss, *SYNTHETIC BIOLOGY. Synthetic communities, the sum of parts.* Science, 2015. **349**(6251): p. 924-5.

84. Heinemann, M. and S. Panke, *Synthetic biology--putting engineering into biology.* Bioinformatics, 2006. **22**(22): p. 2790-9.

85. Bradley, R.W., M. Buck, and B. Wang, *Tools and Principles for Microbial Gene Circuit Engineering.* J Mol Biol, 2016. **428**(5 Pt B): p. 862-88.

86. Cummins, J. and M. Tangney, *Bacteria and tumours: causative agents or opportunistic inhabitants?* Infect Agent Cancer, 2013. **8**(1): p. 11.

87. Baban, C.K., et al., *Bacteria as vectors for gene therapy of cancer.* Bioeng Bugs, 2010. **1**(6): p. 385-94.

88. Forbes, N.S., *Engineering the perfect (bacterial) cancer therapy.* Nat Rev Cancer, 2010. **10**(11): p. 785-94.

89. Urbaniak, C., et al., *The Microbiota of Breast Tissue and Its Association with Breast Cancer.* Appl Environ Microbiol, 2016. **82**(16): p. 5039-48.

90. Urbaniak, C., et al., *Microbiota of human breast tissue.* Appl Environ Microbiol, 2014. **80**(10): p. 3007-14.

91. Xuan, C., et al., *Microbial Dysbiosis Is Associated with Human Breast Cancer.* PLOS ONE, 2014. **9**(1): p. e83744.

92. Banerjee, S., et al., *Distinct microbiological signatures associated with triple negative breast cancer.* Sci Rep, 2015. **5**: p. 15162.

93. Morrissey, D., G.C. O'Sullivan, and M. Tangney, *Tumour targeting with systemically administered bacteria.* Curr Gene Ther, 2010. **10**(1): p. 3-14.

94. Thornlow, D.N., et al., *Persistent enhancement of bacterial motility increases tumor penetration.* Biotechnol Bioeng, 2015. **112**(11): p. 2397-405.

95. Hieken, T.J., et al., *The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease.* Scientific Reports, 2016. **6**: p. 30751.

96. Thompson, K.J., et al., *A comprehensive analysis of breast cancer microbiota and host gene expression.* PLOS ONE, 2017. **12**(11): p. e0188873.

97. Urbaniak, C., et al., *The Microbiota of Breast Tissue and Its Association with Breast Cancer.* Applied and Environmental Microbiology, 2016. **82**(16): p. 5039-5048.

98. Wang, H., et al., *Breast tissue, oral and urinary microbiomes in breast cancer.* Oncotarget, 2017. **8**(50): p. 88122-88138.

99. Fernández, L., et al., *The human milk microbiota: Origin and potential roles in health and disease.* Pharmacological Research, 2013. **69**(1): p. 1-10.

100. Hunt, K.M., et al., *Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk.* PLoS ONE, 2011. **6**(6): p. e21313.

101. Donnet-Hughes, A., et al., *Potential role of the intestinal microbiota of the mother in neonatal immune education.* Proc Nutr Soc, 2010. **69**(3): p. 407-15.

102. Urbaniak, C., J.P. Burton, and G. Reid, *Breast, milk and microbes: a complex relationship that does not end with lactation.* Womens Health (Lond), 2012. **8**(4): p. 385-98.

103. Kim, S.H., et al., *High efficacy of a Listeria-based vaccine against metastatic breast cancer reveals a dual mode of action.* Cancer research, 2009. **69**(14): p. 5860-5866.

104. Macpherson, A.J. and T. Uhr, *Induction of Protective IgA by Intestinal Dendritic Cells Carrying Commensal Bacteria.* Science, 2004. **303**(5664): p. 1662.

105. Rescigno, M., et al., *Dendritic cells express tight junction proteins and penetrate gut epithelial monolayers to sample bacteria.* Nature Immunology, 2001. **2**: p. 361.

106. Chan, A.A., et al., *Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors.* Scientific Reports, 2016. **6**(1): p. 28061.

107. Jiménez, E., et al., *Oral Administration of Lactobacillus Strains Isolated from Breast Milk as an Alternative for the Treatment of Infectious Mastitis during Lactation.* Applied and Environmental Microbiology, 2008. **74**(15): p. 4650-4655.

108. Meyer, K.M., et al., *20: Maternal diet structures the breast milk microbiome in association with human milk oligosaccharides and gut-associated bacteria.* American Journal of Obstetrics & Gynecology, 2017. **216**(1): p. S15.

109. Biffi, A., et al., *Antiproliferative effect of fermented milk on the growth of a human breast cancer cell line.* Nutr Cancer, 1997. **28**(1): p. 93-9.

110. Motevaseli, E., A. Dianatpour, and S. Ghafouri-Fard, *The Role of Probiotics in Cancer Treatment: Emphasis on their In Vivo and In Vitro Anti-metastatic Effects.* International Journal of Molecular and Cellular Medicine, 2017. **6**(2): p. 66-76.

111. Yu, A.-Q. and L. Li, *The Potential Role of Probiotics in Cancer Prevention and Treatment.* Nutrition and Cancer, 2016. **68**(4): p. 535-544.

112. de Moreno de LeBlanc, A., et al., *Effects of milk fermented by Lactobacillus helveticus R389 on immune cells associated to mammary glands in normal and a breast cancer model.* Immunobiology, 2005. **210**(5): p. 349-58.

113. Kosaka, A., et al., *Lactococcus lactis subsp. cremoris FC triggers IFN-gamma production from NK and T cells via IL-12 and IL-18.* Int Immunopharmacol, 2012. **14**(4): p. 729-33.

114. Plottel, Claudia S. and Martin J. Blaser, *Microbiome and Malignancy.* Cell Host & Microbe, 2011. **10**(4): p. 324-335.

115. Kwa, M., et al., *The Intestinal Microbiome and Estrogen Receptor-Positive Female Breast Cancer.* Journal of the National Cancer Institute, 2016. **108**(8): p. djw029.

116. Gaya, P., et al., *Phytoestrogen Metabolism by Adult Human Gut Microbiota.* Molecules, 2016. **21**(8).

117. Pollock, C.B., et al., *Strigolactones: a novel class of phytohormones that inhibit the growth and survival of breast cancer cells and breast cancer stem-like enriched mammosphere cells.* Breast Cancer Research and Treatment, 2012. **134**(3): p. 1041-1055.

118. Velicer, C.M., et al., *Antibiotic use in relation to the risk of breast cancer.* JAMA, 2004. **291**(7): p. 827-835.

119. Mezouar, S., et al., *Microbiome and the immune system: From a healthy steady-state to allergy associated disruption.* Human Microbiome Journal, 2018. **10**: p. 11-20.

120. Belkaid, Y. and T.W. Hand, *Role of the microbiota in immunity and inflammation.* Cell, 2014. **157**(1): p. 121-141.

121. Gorjifard, S. and R.S. Goldszmid, *Microbiota—myeloid cell crosstalk beyond the gut.* Journal of Leukocyte Biology, 2016. **100**(5): p. 865-879.

122. Fessler, J., V. Matson, and T.F. Gajewski, *Exploring the emerging role of the microbiome in cancer immunotherapy.* Journal for ImmunoTherapy of Cancer, 2019. **7**(1): p. 108.

123. Ivanov, I.I., et al., *Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine.* Cell host & microbe, 2008. **4**(4): p. 337-349.

124. Mazmanian, S.K., et al., *An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System.* Cell, 2005. **122**(1): p. 107-118.

125. Li, W., et al., *Gut microbiome and cancer immunotherapy.* Cancer Letters, 2019. **447**: p. 41-47.

126. Temraz, S., et al., *Gut Microbiome: A Promising Biomarker for Immunotherapy in Colorectal Cancer.* International journal of molecular sciences, 2019. **20**(17): p. 4155.

127. Shui, L., et al., *Gut Microbiome as a Potential Factor for Modulating Resistance to Cancer Immunotherapy.* Frontiers in Immunology, 2020. **10**(2989).

128. Sivan, A., et al., *Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy.* Science (New York, N.Y.), 2015. **350**(6264): p. 1084-1089.

129. Frankel, A.E., et al., *Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients.* Neoplasia, 2017. **19**(10): p. 848-855.

130. Matson, V., et al., *The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients.* Science, 2018. **359**(6371): p. 104-108.

131. Routy, B., et al., *Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors.* Science, 2018. **359**(6371): p. 91-97.

132. Vetizou, M., et al., *Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota.* Science, 2015. **350**(6264): p. 1079-84.

133. Ahmad, S., et al., *Induction of effective antitumor response after mucosal bacterial vector mediated DNA vaccination with endogenous prostate cancer specific antigen.* J Urol, 2011. **186**(2): p. 687-93.

134. Wood, L.M. and Y. Paterson, *Attenuated Listeria monocytogenes: a powerful and versatile vector for the future of tumor immunotherapy.* Front Cell Infect Microbiol, 2014. **4**: p. 51.

135. Low, K.B., et al., *Construction of VNP20009: a novel, genetically stable antibiotic-sensitive strain of tumor-targeting Salmonella for parenteral administration in humans.* Methods Mol Med, 2004. **90**: p. 47-60.

136. Pizarro-Cerda, J. and K. Tedin, *The bacterial signal molecule, ppGpp, regulates Salmonella virulence gene expression.* Mol Microbiol, 2004. **52**(6): p. 1827-44.

137. Stritzker, J., et al., *Myristoylation negative msbB-mutants of probiotic E. coli Nissle 1917 retain tumor specific colonization properties but show less side effects in immunocompetent mice.* Bioeng Bugs, 2010. **1**(2): p. 139-45.

138. Yu, B., et al., *Explicit hypoxia targeting with tumor suppression by creating an "obligate" anaerobic Salmonella Typhimurium strain.* Sci Rep, 2012. **2**: p. 436.

139. Park, S.H., et al., *RGD Peptide Cell-Surface Display Enhances the Targeting and Therapeutic Efficacy of Attenuated Salmonella-mediated Cancer Therapy.* Theranostics, 2016. **6**(10): p. 1672-82.

140. Byrne, W.L., et al., *Bacterial-mediated DNA delivery to tumour associated phagocytic cells.* J Control Release, 2014. **196**: p. 384-93.

141. Anderson, J.C., et al., *Environmentally Controlled Invasion of Cancer Cells by Engineered Bacteria.* Journal of Molecular Biology, 2006. **355**(4): p. 619-627.

142. van Pijkeren, J.P., et al., *A novel Listeria monocytogenes-based DNA delivery system for cancer gene therapy.* Hum Gene Ther, 2010. **21**(4): p. 405-16.

143. Wang, B., M. Barahona, and M. Buck, *Engineering modular and tunable genetic amplifiers for scaling transcriptional signals in cascaded gene networks.* Nucleic Acids Res, 2014. **42**(14): p. 9484-92.

144. Wang, B., M. Barahona, and M. Buck, *A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals.* Biosens Bioelectron, 2013. **40**(1): p. 368-76.

145. MacDonald, J.T., et al., *Computational design approaches and tools for synthetic biology.* Integr Biol (Camb), 2011. **3**(2): p. 97-108.

146. Marchisio, M.A. and J. Stelling, *Computational design tools for synthetic biology.* Curr Opin Biotechnol, 2009. **20**(4): p. 479-85.

147. Nuyts, S., et al., *Radio-responsive recA promoter significantly increases TNFalpha production in recombinant clostridia after 2 Gy irradiation.* Gene Ther, 2001. **8**(15): p. 1197-201.

148. Nguyen, V.H., et al., *Genetically engineered Salmonella typhimurium as an imageable therapeutic probe for cancer.* Cancer Res, 2010. **70**(1): p. 18-23.

149. Ryan, R.M., et al., *Bacterial delivery of a novel cytolysin to hypoxic areas of solid tumors.* Gene Ther, 2009. **16**(3): p. 329-39.

150. Camacho, E.M., et al., *Engineering Salmonella as intracellular factory for effective killing of tumour cells.* Sci Rep, 2016. **6**: p. 30591.

151. Baeshen, M.N., et al., *Production of Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives.* J Microbiol Biotechnol, 2015. **25**(7): p. 953-62.

152. Huang, C.J., H. Lin, and X. Yang, *Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements.* J Ind Microbiol Biotechnol, 2012. **39**(3): p. 383-99.

153. Bermudez-Humaran, L.G., et al., *Engineering lactococci and lactobacilli for human health.* Curr Opin Microbiol, 2013. **16**(3): p. 278-83.

154. Felgner, S., et al., *Bacteria in Cancer Therapy: Renaissance of an Old Concept.* Int J Microbiol, 2016. **2016**: p. 8451728.

155. Gupta, S.K. and P. Shukla, *Advanced technologies for improved expression of recombinant proteins in bacteria: perspectives and applications.* Crit Rev Biotechnol, 2016. **36**(6): p. 1089-1098.

156. Takiishi, T., et al., *Reversal of Diabetes in NOD Mice by Clinical-Grade Proinsulin and IL-10-Secreting Lactococcus lactis in Combination With Low-Dose Anti-CD3 Depends on the Induction of Foxp3-Positive T Cells.* Diabetes, 2017. **66**(2): p. 448-459.

157. Maurer, J., J. Jose, and T.F. Meyer, *Autodisplay: one-component system for efficient surface display and release of soluble recombinant proteins from Escherichia coli.* J Bacteriol, 1997. **179**(3): p. 794-804.

158. Michon, C., et al., *Display of recombinant proteins at the surface of lactic acid bacteria: strategies and applications.* Microb Cell Fact, 2016. **15**: p. 70.

159. Manuel, E.R., et al., *Enhancement of cancer vaccine therapy by systemic delivery of a tumor-targeting Salmonella-based STAT3 shRNA suppresses the growth of established melanoma tumors.* Cancer Res, 2011. **71**(12): p. 4183-91.

160. Tian, Y., et al., *Targeted therapy via oral administration of attenuated Salmonella expression plasmid-vectored Stat3-shRNA cures orthotopically transplanted mouse HCC.* Cancer Gene Ther, 2012. **19**(6): p. 393-401.

161. Lehouritis, P., C. Springer, and M. Tangney, *Bacterial-directed enzyme prodrug therapy.* J Control Release, 2013. **170**(1): p. 120-31.

162. Lehouritis, P., et al., *Activation of multiple chemotherapeutic prodrugs by the natural enzymolome of tumour-localised probiotic bacteria.* J Control Release, 2016. **222**: p. 9-17.

163. Rajendran, S., et al., *Targeting of breast metastases using a viral gene vector with tumour-selective transcription.* Anticancer Res, 2011. **31**(5): p. 1627-35.

164. Danino, T., et al., *Programmable probiotics for detection of cancer in urine.* Sci Transl Med, 2015. **7**(289): p. 289ra84.

165. Panteli, J.T., et al., *Genetically modified bacteria as a tool to detect microscopic solid tumor masses with triggered release of a recombinant biomarker.* Integr Biol (Camb), 2015. **7**(4): p. 423-34.

166. Bryant, L.M., et al., *Lessons learned from the clinical development and market authorization of Glybera.* Hum Gene Ther Clin Dev, 2013. **24**(2): p. 55-64.

167. Lichty, B.D., et al., *Going viral with cancer immunotherapy.* Nat Rev Cancer, 2014. **14**(8): p. 559-67.

168. Le, D.T., et al., *Safety and survival with GVAX pancreas prime and Listeria Monocytogenes-expressing mesothelin (CRS-207) boost vaccines for metastatic pancreatic cancer.* J Clin Oncol, 2015. **33**(12): p. 1325-33.

169. Brockstedt, D.G., et al., *Listeria-based cancer vaccines that segregate immunogenicity from toxicity.* Proc Natl Acad Sci U S A, 2004. **101**(38): p. 13832-7.

170. Verch, T., Z.K. Pan, and Y. Paterson, *Listeria monocytogenes-based antibiotic resistance gene-free antigen delivery system applicable to other bacterial vectors and DNA vaccines.* Infect Immun, 2004. **72**(11): p. 6418-25.

171. Gaffney, E.F., et al., *Factors that drive the increasing use of FFPE tissue in basic and translational cancer research.* Biotechnic & Histochemistry, 2018. **93**(5): p. 373-386.

172. van Beers, E.H., et al., *A multiplex PCR predictor for aCGH success of FFPE samples.* British journal of cancer, 2006. **94**(2): p. 333-337.

173. Blow, N., *Tissue issues.* Nature, 2007. **448**(7156): p. 959-960.

174. Mathieson, W. and G. Thomas, *Using FFPE Tissue in Genomic Analyses: Advantages, Disadvantages and the Role of Biospecimen Science.* Current Pathobiology Reports, 2019. **7**(3): p. 35-40.

175. Kerick, M., et al., *Targeted high throughput sequencing in clinical cancer Settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity.* BMC Medical Genomics, 2011. **4**(1): p. 68.

176. Hedegaard, J., et al., *Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue.* PloS one, 2014. **9**(5): p. e98187-e98187.

177. McDonough, S.J., et al., *Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods.* PloS one, 2019. **14**(4): p. e0211400-e0211400.

178. Kresse, S.H., et al., *Evaluation of commercial DNA and RNA extraction methods for high-throughput sequencing of FFPE samples.* PLOS ONE, 2018. **13**(5): p. e0197456.

179. Siebolts, U., et al., *Tissues from routine pathology archives are suitable for microRNA analyses by quantitative PCR.* Journal of Clinical Pathology, 2009. **62**(1): p. 84-88.

180. Fanelli, M., et al., *Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue.* Nature Protocols, 2011. **6**(12): p. 1905-1919.

181. Zhang, P., et al., *The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies.* International Journal of Genomics, 2017. **2017**: p. 9.

182. Srinivasan, M., D. Sedmak, and S. Jewell, *Effect of fixatives and tissue processing on the content and integrity of nucleic acids.* The American journal of pathology, 2002. **161**(6): p. 1961-1971.

183. Vitosevic, K., et al., *Effect of formalin fixation on pcr amplification of DNA isolated from healthy autopsy tissues.* Acta Histochem, 2018. **120**(8): p. 780-788.

184. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* Journal of Experimental Biology, 2007. **210**(9): p. 1518-1525.

185. Forbes, J.D., et al., *Metagenomics: The Next Culture-Independent Game Changer.* Frontiers in Microbiology, 2017. **8**(1069).

186. Knight, R., et al., *Best practices for analysing microbiomes.* Nature Reviews Microbiology, 2018. **16**(7): p. 410-422.

187. Proctor, L., et al., *A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016.* Microbiome, 2019. **7**(1): p. 31.

188. Ding, T. and P.D. Schloss, *Dynamics and associations of microbial community types across the human body.* Nature, 2014. **509**(7500): p. 357-360.

189. Cho, I. and M.J. Blaser, *The human microbiome: at the interface of health and disease.* Nature Reviews Genetics, 2012. **13**(4): p. 260-270.

190. Pannaraj, P.S., et al., *Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome.* JAMA Pediatrics, 2017. **171**(7): p. 647-654.

191. Castillo, D.J., et al., *The Healthy Human Blood Microbiome: Fact or Fiction?* Frontiers in Cellular and Infection Microbiology, 2019. **9**(148).

192. Stinson, L.F., et al., *The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth.* Frontiers in Microbiology, 2019. **10**(1124).

193. Ozkan, J., et al., *Identification and Visualization of a Distinct Microbiome in Ocular Surface Conjunctival Tissue.* Investigative Ophthalmology & Visual Science, 2018. **59**(10): p. 4268-4276.

194. Zhou, B., et al., *The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients.* Scientific Reports, 2019. **9**(1): p. 1691.

195. Chen, J., et al., *The microbiome and breast cancer: a review.* Breast Cancer Res Treat, 2019.

196. Beck, J.M., V.B. Young, and G.B. Huffnagle, *The microbiome of the lung.* Translational research : the journal of laboratory and clinical medicine, 2012. **160**(4): p. 258-266.

197. Huffnagle, G.B., R.P. Dickson, and N.W. Lukacs, *The respiratory tract microbiome and lung inflammation: a two-way street.* Mucosal Immunology, 2017. **10**(2): p. 299-306.

198. Marsh, R.L., et al., *The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx.* Microbiome, 2016. **4**(1): p. 37.

199. Emery, D.C., et al., *16S rRNA Next Generation Sequencing Analysis Shows Bacteria in Alzheimer's Post-Mortem Brain.* Frontiers in Aging Neuroscience, 2017. **9**(195).

200. Stewart, C.J., et al., *Using formalin fixed paraffin embedded tissue to characterize the preterm gut microbiota in necrotising enterocolitis and spontaneous isolated perforation using marginal and diseased tissue.* BMC Microbiology, 2019. **19**(1): p. 52.

201. Hart, J.D., et al., *16S rRNA sequencing in molecular microbiological diagnosis of bacterial infections in the autopsy setting.* Pathology, 2014. **46**: p. S113.

202. Racsa, L.D., et al., *Identification of bacterial pathogens from formalin-fixed, paraffin-embedded tissues by using 16S sequencing: retrospective correlation of results to clinicians' responses.* Human Pathology, 2017. **59**: p. 132-138.

203. Banerjee, S., et al., *Distinct Microbial Signatures Associated With Different Breast Cancer Types.* Frontiers in Microbiology, 2018. **9**(951).

204. Baldwin, D.A., et al., *Metagenomic assay for identification of microbial pathogens in tumor tissues.* mBio, 2014. **5**(5): p. e01714.

205. Riquelme, E., et al., *Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes.* Cell, 2019. **178**(4): p. 795-806.e12.

206. Einaga, N., et al., *Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.* PLOS ONE, 2017. **12**(5): p. e0176280.

207. Bailey, S.T., et al., *High-quality whole-genome sequencing of FFPE samples.* Journal of Clinical Oncology, 2018. **36**(15_suppl): p. e13500-e13500.

208. Bettoni, F., et al., *A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology.* Exp Mol Pathol, 2017. **103**(3): p. 294-299.

209. Lindahl, T. and A. Andersson, *Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid.* Biochemistry, 1972. **11**(19): p. 3618-3623.

210. Lindahl, T. and B. Nyberg, *Rate of depurination of native deoxyribonucleic acid.* Biochemistry, 1972. **11**(19): p. 3610-3618.

211. Feldman, M.Y., *Reactions of nucleic acids and nucleoproteins with formaldehyde.* Prog Nucleic Acid Res Mol Biol, 1973. **13**: p. 1-49.

212. Siomin, Y.A., V.V. Simonov, and A.M. Poverenny, *The reaction of formaldehyde with deoxynucleotides and DNA in the presence of amino acids and lysine-rich histone.* Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis, 1973. **331**(1): p. 27-32.

213. McGhee, J.D. and P.H. Von Hippel, *Formaldehyde as a probe of DNA structure. I. Reaction with exocyclic amino groups of DNA bases.* Biochemistry, 1975. **14**(6): p. 1281-1296.

214. McGhee, J.D. and P.H. Von Hippel, *Formaldehyde as a probe of DNA structure. II. Reaction with endocyclic imino groups of DNA bases.* Biochemistry, 1975. **14**(6): p. 1297-1303.

215. McGhee, J.D. and P.H. Von Hippel, *Formaldehyde as a probe of DNA structure. 4. Mechanism of the initial reaction of formaldehyde with DNA.* Biochemistry, 1977. **16**(15): p. 3276-3293.

216. McGhee, J.D. and P.H. von Hippel, *Formaldehyde as a probe of DNA structure. 3. Equilibrium denaturation of DNA and synthetic polynucleotides.* Biochemistry, 1977. **16**(15): p. 3267-76.

217. Chaw, Y.F.M., et al., *Isolation and identification of cross-links from formaldehyde-treated nucleic acids.* Biochemistry, 1980. **19**(24): p. 5525-5531.

218. Huang, H. and P.B. Hopkins, *DNA interstrand cross-linking by formaldehyde: nucleotide sequence preference and covalent structure of the predominant cross-link formed in synthetic oligonucleotides.* Journal of the American Chemical Society, 1993. **115**(21): p. 9402-9408.

219. Kamps, J.J.A.G., et al., *How formaldehyde reacts with amino acids.* Communications Chemistry, 2019. **2**(1): p. 126.

220. Haile, S., et al., *Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples.* Nucleic acids research, 2019. **47**(2): p. e12-e12.

221. Shishodia, S., et al., *NMR analyses on N-hydroxymethylated nucleobases – implications for formaldehyde toxicity and nucleic acid demethylases.* Organic & Biomolecular Chemistry, 2018. **16**(21): p. 4021-4032.

222. Robbe, P., et al., *Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.* Genetics in Medicine, 2018. **20**(10): p. 1196-1205.

223. Groelz, D., et al., *Impact of storage conditions on the quality of nucleic acids in paraffin embedded tissues.* PLOS ONE, 2018. **13**(9): p. e0203608.

224. Yang, Z., et al., *Interstrand cross-links arising from strand breaks at true abasic sites in duplex DNA.* Nucleic Acids Res, 2017. **45**(11): p. 6275-6283.

225. Lu, K., et al., *Use of LC-MS/MS and Stable Isotopes to Differentiate Hydroxymethyl and Methyl DNA Adducts from Formaldehyde and Nitrosodimethylamine.* Chemical Research in Toxicology, 2012. **25**(3): p. 664-675.

226. Nawy, T., *DNA variants or DNA damage?* Nature Methods, 2017. **14**(4): p. 341-341.

227. Chen, L., et al., *DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification.* Science, 2017. **355**(6326): p. 752.

228. Do, H. and A. Dobrovic, *Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization.* Clinical Chemistry, 2015. **61**(1): p. 64-71.

229. Hoffman, E.A., et al., *Formaldehyde crosslinking: a tool for the study of chromatin complexes.* J Biol Chem, 2015. **290**(44): p. 26404-11.

230. Lu, K., et al., *Structural characterization of formaldehyde-induced cross-links between amino acids and deoxynucleosides and their oligomers.* J Am Chem Soc, 2010. **132**(10): p. 3388-99.

231. Chang, Y.-T. and G.H. Loew, *Reaction Mechanisms of Formaldehyde with Endocyclic Imino Groups of Nucleic Acid Bases.* Journal of the American Chemical Society, 1994. **116**(8): p. 3548-3555.

232. Huang, H., M.S. Solomon, and P.B. Hopkins, *Formaldehyde preferentially interstrand cross-links duplex DNA through deoxyadenosine residues at the sequence 5'-d(AT).* Journal of the American Chemical Society, 1992. **114**(23): p. 9240-9241.

233. Metz, B., et al., *Identification of Formaldehyde-induced Modifications in Proteins: REACTIONS WITH MODEL PEPTIDES.* Journal of Biological Chemistry, 2004. **279**(8): p. 6235-6243.

234. Toews, J., et al., *Mass spectrometric identification of formaldehyde-induced peptide modifications under in vivo protein cross-linking conditions.* Analytica Chimica Acta, 2008. **618**(2): p. 168-183.

235. Metz, B., et al., *Identification of Formaldehyde-Induced Modifications in Proteins: Reactions with Insulin.* Bioconjugate Chemistry, 2006. **17**(3): p. 815-822.

236. Toews, J., J.C. Rogalski, and J. Kast, *Accessibility governs the relative reactivity of basic residues in formaldehyde-induced protein modifications.* Analytica Chimica Acta, 2010. **676**(1): p. 60-67.

237. Schmiedeberg, L., et al., *A Temporal Threshold for Formaldehyde Crosslinking and Fixation.* PLOS ONE, 2009. **4**(2): p. e4636.

238. Solomon, M.J. and A. Varshavsky, *Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures.* Proc Natl Acad Sci U S A, 1985. **82**(19): p. 6470-4.

239. Lu, K., et al., *Formaldehyde-Induced Histone Modifications in Vitro.* Chemical Research in Toxicology, 2008. **21**(8): p. 1586-1593.

240. Tretyakova, N.Y., A. Groehler, and S. Ji, *DNA–Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes.* Accounts of Chemical Research, 2015. **48**(6): p. 1631-1644.

241. Dutta, S., G. Chowdhury, and K.S. Gates, *Interstrand Cross-Links Generated by Abasic Sites in Duplex DNA.* Journal of the American Chemical Society, 2007. **129**(7): p. 1852-1853.

242. Kennedy-Darling, J. and L.M. Smith, *Measuring the Formaldehyde Protein–DNA Cross-Link Reversal Rate.* Analytical Chemistry, 2014. **86**(12): p. 5678-5681.

243. Rait, V.K., et al., *Conversions of formaldehyde-modified 2'-deoxyadenosine 5'-monophosphate in conditions modeling formalin-fixed tissue dehydration.* The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society, 2006. **54**(3): p. 301-310.

244. Bass, B.P., et al., *A review of preanalytical factors affecting molecular, protein, and morphological analysis of formalin-fixed, paraffin-embedded*

(FFPE) tissue: how well do you know your FFPE specimen? Arch Pathol Lab Med, 2014. **138**(11): p. 1520-30.

245. Evers, D.L., et al., *Paraffin Embedding Contributes to RNA Aggregation, Reduced RNA Yield, and Low RNA Quality.* The Journal of Molecular Diagnostics, 2011. **13**(6): p. 687-694.

246. Xie, R., et al., *Factors Influencing the Degradation of Archival Formalin-Fixed Paraffin-Embedded Tissue Sections.* Journal of Histochemistry & Cytochemistry, 2011. **59**(4): p. 356-365.

247. Watanabe, M., et al., *Estimation of age-related DNA degradation from formalin-fixed and paraffin-embedded tissue according to the extraction methods.* Experimental and therapeutic medicine, 2017. **14**(3): p. 2683-2688.

248. Yudkina, A.V., A.P. Dvornikova, and D.O. Zharkov, *Variable termination sites of DNA polymerases encountering a DNA–protein cross-link.* PLOS ONE, 2018. **13**(6): p. e0198480.

249. Kawashima, Y., et al., *Efficient extraction of proteins from formalin-fixed paraffin-embedded tissues requires higher concentration of tris(hydroxymethyl)aminomethane.* Clinical proteomics, 2014. **11**(1): p. 4-4.

250. Einaga, N., et al., *Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.* PloS one, 2017. **12**(5): p. e0176280-e0176280.

251. McDonough, S.J., et al., *Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods.* 2019. **14**(4): p. e0211400.

252. Winogradoff, D., S. John, and A. Aksimentiev, *Protein unfolding by SDS: the microscopic mechanisms and the properties of the SDS-protein assembly.* Nanoscale, 2020. **12**(9): p. 5422-5434.

253. Bhattacharya, S. and S.S. Mandal, *Interaction of surfactants with DNA. Role of hydrophobicity and surface charge on intercalation and DNA melting.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 1997. **1323**(1): p. 29-44.

254. Huijsmans, C.J., et al., *Comparative analysis of four methods to extract DNA from paraffin-embedded tissues: effect on downstream molecular applications.* BMC research notes, 2010. **3**: p. 239-239.

255. Wingfield, P.T., *Use of protein folding reagents.* Current protocols in protein science, 2001. **Appendix 3**: p. Appendix-3A.

256. Kubíčková, A., et al., *Guanidinium Cations Pair with Positively Charged Arginine Side Chains in Water.* The Journal of Physical Chemistry Letters, 2011. **2**(12): p. 1387-1389.

257. Mason, P.E., et al., *The Structure of Aqueous Guanidinium Chloride Solutions.* Journal of the American Chemical Society, 2004. **126**(37): p. 11462-11470.

258. Chein, Y.-H. and N. Davidson, *RNA:DNA hybrids are more stable than DNA:DNA duplexes in concentrated perchlorate and trichloroacetate solutions.* Nucleic Acids Research, 1978. **5**(5): p. 1627-1637.

259. Lambert, D. and D.E. Draper, *Denaturation of RNA secondary and tertiary structure by urea: simple unfolded state models and free energy parameters account for measured m-values.* Biochemistry, 2012. **51**(44): p. 9014-9026.

260. Ness, J.V. and L. Chen, *The use of oliodeoxynucleotide probes in chaotrope-based hybridization solutions.* Nucleic Acids Research, 1991. **19**(19): p. 5143-5151.

261. Yang, H.J. and C.L. Tsou, *Inactivation during denaturation of ribonuclease A by guanidinium chloride is accompanied by unfolding at the active site.* The Biochemical journal, 1995. **305 (Pt 2)**(Pt 2): p. 379-384.

262. Poulsen, J.W., et al., *Using Guanidine-Hydrochloride for Fast and Efficient Protein Digestion and Single-step Affinity-purification Mass Spectrometry.* Journal of Proteome Research, 2013. **12**(2): p. 1020-1030.

263. Jain, N., et al., *Direct Observation of the Intrinsic Backbone Torsional Mobility of Disordered Proteins.* Biophysical journal, 2016. **111**(4): p. 768-774.

264. Marcus, Y., *The guanidinium ion.* The Journal of Chemical Thermodynamics, 2012. **48**: p. 70-74.

265. Hamoud, F., et al., *Guanidine hydrochloride aminomethylation with formaldehyde and simplest amino acids.* Russian Journal of Organic Chemistry, 2017. **53**(8): p. 1258-1267.

266. Lindahl, T., *Instability and decay of the primary structure of DNA.* Nature, 1993. **362**(6422): p. 709-715.

267. Bonnet, J., et al., *Chain and conformation stability of solid-state DNA: implications for room temperature storage.* Nucleic acids research, 2010. **38**(5): p. 1531-1546.

268. Yoshioka, S. and Y. Aso, *Correlations between molecular mobility and chemical stability during storage of amorphous pharmaceuticals.* J Pharm Sci, 2007. **96**(5): p. 960-81.

269. Vesnaver, G., et al., *Influence of abasic and anucleosidic sites on the stability, conformation, and melting behavior of a DNA duplex: correlations of thermodynamic and structural data.* Proceedings of the National Academy of Sciences, 1989. **86**(10): p. 3614-3618.

270. Gates, K.S., *An Overview of Chemical Processes That Damage Cellular DNA: Spontaneous Hydrolysis, Alkylation, and Reactions with Radicals.* Chemical Research in Toxicology, 2009. **22**(11): p. 1747-1760.

271. Sugiyama, H., et al., *Chemistry of Thermal Degradation of Abasic Sites in DNA. Mechanistic Investigation on Thermal DNA Strand Cleavage of Alkylated DNA.* Chemical Research in Toxicology, 1994. **7**(5): p. 673-683.

272. Sikorsky, J.A., et al., *Effect of DNA damage on PCR amplification efficiency with the relative threshold cycle method.* Biochem Biophys Res Commun, 2004. **323**(3): p. 823-30.

273. Sikorsky, J.A., et al., *DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency.* Biochemical and Biophysical Research Communications, 2007. **355**(2): p. 431-437.

274. Shibutani, S., M. Takeshita, and A.P. Grollman, *Translesional Synthesis on DNA Templates Containing a Single Abasic Site: A MECHANISTIC STUDY OF THE "A RULE".* Journal of Biological Chemistry, 1997. **272**(21): p. 13916-13922.

275. Heyn, P., et al., *Road blocks on paleogenomes--polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA.* Nucleic Acids Res, 2010. **38**(16): p. e161.

276. Guillet, M. and S. Boiteux, *Origin of endogenous DNA abasic sites in Saccharomyces cerevisiae.* Molecular and cellular biology, 2003. **23**(22): p. 8386-8394.

277. Dietrich, D., et al., *Improved PCR performance using template DNA from formalin-fixed and paraffin-embedded tissues by overcoming PCR inhibition.* PloS one, 2013. **8**(10): p. e77771-e77771.

278. Zimmermann, J., et al., *DNA damage in preserved specimens and tissue samples: a molecular assessment.* Frontiers in Zoology, 2008. **5**(1): p. 18.

279. Hegde, M.L., T.K. Hazra, and S. Mitra, *Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells.* Cell research, 2008. **18**(1): p. 27-47.

280. Davis, L., Y. Zhang, and N. Maizels, *Assaying Repair at DNA Nicks.* Methods in enzymology, 2018. **601**: p. 71-89.

281. Williams, C., et al., *A high frequency of sequence alterations is due to formalin fixation of archival specimens.* Am J Pathol, 1999. **155**(5): p. 1467-71.

282. Do, H. and A. Dobrovic, *Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase.* Oncotarget, 2012. **3**(5): p. 546-58.

283. Bettoni, F., et al., *A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology.* Experimental and Molecular Pathology, 2017. **103**(3): p. 294-299.

284. McDonough, S.J., et al., *Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods.* PLOS ONE, 2019. **14**(4): p. e0211400.

285. Bonnet, E., et al., *Performance comparison of three DNA extraction kits on human whole-exome data from formalin-fixed paraffin-embedded normal and tumor samples.* PloS one, 2018. **13**(4): p. e0195471-e0195471.

286. Spencer, D.H., et al., *Comparison of Clinical Targeted Next-Generation Sequence Data from Formalin-Fixed and Fresh-Frozen Tissue Specimens.* The Journal of Molecular Diagnostics, 2013. **15**(5): p. 623-633.

287. Kim, S., et al., *Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine.* J Mol Diagn, 2017. **19**(1): p. 137-146.

288. Schweiger, M.R., et al., *Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis.* PloS one, 2009. **4**(5): p. e5548-e5548.

289. Robbe, P., et al., *Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.* Genet Med, 2018. **20**(10): p. 1196-1205.

290. Wong, S.Q., et al., *Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing.* BMC Medical Genomics, 2014. **7**(1): p. 23.

291. Carrick, D.M., et al., *Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue.* PLOS ONE, 2015. **10**(7): p. e0127353.

292. Lindahl, T., *DNA glycosylases, endonucleases for apurinic/apyrimidinic sites, and base excision-repair.* Prog Nucleic Acid Res Mol Biol, 1979. **22**: p. 135-92.

293. Krokan, H.E., F. Drabløs, and G. Slupphaug, *Uracil in DNA – occurrence, consequences and repair.* Oncogene, 2002. **21**(58): p. 8935-8948.

294. Krokan, H.E. and M. Bjørås, *Base Excision Repair.* Cold Spring Harbor Perspectives in Biology, 2013. **5**(4).

295. Bellacosa, A. and A.C. Drohat, *Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites.* DNA repair, 2015. **32**: p. 33-42.

296. Poole, A., D. Penny, and B.-M. Sjöberg, *Confounded cytosine! Tinkering and the evolution of DNA.* Nature Reviews Molecular Cell Biology, 2001. **2**(2): p. 147-151.

297. Schomacher, L. and C. Niehrs, *DNA repair and erasure of 5-methylcytosine in vertebrates.* Bioessays, 2017. **39**(3).

298. Peluso, M.E.M., et al., *Oxidative DNA damage and formalin-fixation procedures.* Toxicology Research, 2014. **3**(5): p. 341-349.

299. Bono, R., et al., *Malondialdehyde-deoxyguanosine adduct formation in workers of pathology wards: the role of air formaldehyde exposure.* Chem Res Toxicol, 2010. **23**(8): p. 1342-8.

300. Wallace, S.S., *Base excision repair: a critical player in many games.* DNA repair, 2014. **19**: p. 14-26.

301. Hegde, M.L., T. Izumi, and S. Mitra, *Oxidized base damage and single-strand break repair in mammalian genomes: role of disordered regions and posttranslational modifications in early enzymes.* Progress in molecular biology and translational science, 2012. **110**: p. 123-153.

302. Munchel, S., et al., *Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics.* Oncotarget, 2015. **6**(28): p. 25943-25961.

303. Oh, E., et al., *Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples.* PLOS ONE, 2015. **10**(12): p. e0144162.

304. Guo, M., et al., *Quality and concordance of genotyping array data of 12,064 samples from 5840 cancer patients.* Genomics, 2019. **111**(4): p. 950-957.

305. O'Brien, C.L., et al., *5. Molecular profiling of bacterial DNA isolated from formalin-fixed paraffin-embedded (FFPE) tissue: a comparative study.* Pathology, 2011. **43**: p. S91.

306. Thanbichler, M., P.H. Viollier, and L. Shapiro, *The structure and function of the bacterial chromosome.* Current Opinion in Genetics & Development, 2005. **15**(2): p. 153-162.

307. Webb, C.D., et al., *Bipolar Localization of the Replication Origin Regions of Chromosomes in Vegetative and Sporulating Cells of B. subtilis.* Cell, 1997. **88**(5): p. 667-674.

308. Kow, Y.W., *Repair of deaminated bases in DNA12 1Guest Editor: Miral Dizdaroglu 2This article is part of a series of reviews on "Oxidative DNA Damage and Repair." The full list of papers may be found on the homepage of the journal.* Free Radical Biology and Medicine, 2002. **33**(7): p. 886-893.

309. Frickmann, H., et al., *Next-generation sequencing for hypothesis-free genomic detection of invasive tropical infections in poly-microbially contaminated, formalin-fixed, paraffin-embedded tissue samples – a proof-of-principle assessment.* BMC Microbiology, 2019. **19**(1): p. 75.

310. Chen, L., et al., *DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification.* Science, 2017. **355**(6326): p. 752-756.

311. Krwawicz, J., et al., *Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease.* Acta Biochim Pol, 2007. **54**(3): p. 413-34.

312. Dalhus, B., et al., *DNA base repair – recognition and initiation of catalysis.* FEMS Microbiology Reviews, 2009. **33**(6): p. 1044-1078.

313. Jacobs, A.L. and P. Schär, *DNA glycosylases: in DNA repair and beyond.* Chromosoma, 2012. **121**(1): p. 1-20.

314. Dizdaroglu, M., E. Coskun, and P. Jaruga, *Repair of oxidatively induced DNA damage by DNA glycosylases: Mechanisms of action, substrate specificities and excision kinetics.* Mutation Research/Reviews in Mutation Research, 2017. **771**: p. 99-127.

CHAPTER 2:

Protoblock - A biological standard for formalin fixed samples

ABSTRACT

Formalin-fixed, paraffin-embedded (FFPE) tissue samples are being recognised as viable source material for bacterial analysis. However, several features of this sample type have limited their use for microbiome research. Among these, the lack of standardise methods or workflows. Now, the development of such workflows, could be facilitated by biological standards. In fact, the development and systematic use of reliable standards has been set as a key priority for microbiome research. As such, we aimed at developing a standard for the microbiome analysis of FFPE sample, namely, the Protoblock

The Protoblock is a cell matrix, which can be populated with cell types and numbers as desired, such as to resemble those of the FFPE tissue specimens. Its accuracy for representing bacterial load and cell architecture proven by microscopy. With this model, the performance of the human gold standard FFPE kit for microbiome analysis of FFPE samples was evaluated, and found unsuitable for microbiome research. Additionally, the Protoblock allowed for the characterisation of bacterial FFPE DNA, where it was found highly fragmented ($\bar{x}$ length = 143 bp), a poor PCR template (with a log-fold loss of amplifiable 200 bp fragments) and with a significant extent of sequence alterations. Finally, this model also allowed for the characterisation of FFPE contaminants. Evidence raised here indicates that to unlock the potential of FFPE samples for microbiome analysis, it is necessary to develop a robust quality control system. The Protoblock presented in this study is foundational in building towards this.

INTRODUCTION

Increased sequencing capabilities have driven progress in the study of the human microbiome [1-3], and distinct microbial profiles have been reported in body sites previously thought of as sterile (although many are potentially influenced by environmental contamination) [4-9]. These discoveries have steered a higher demand for patient samples, availability of which can be highly constrained when sampling from body sites that involve invasive sampling procedures [10, 11].

In an attempt to satisfy this demand, the use of formalin fixed paraffin embedded tissue (FFPE) has been explored for microbiome research [12-19]. FFPE tissue is the gold standard for pathology tissue storage and thus represents the largest collection of available patient material [20-22]. The availability of this material has been vital for progress in human genomics and numerous sequencing workflows have been designed to enable use of, or are based upon, these samples [23-28]. The use of FFPE tissue for microbiome research could open access to large sample cohorts (guaranteeing statistical power), accompanied by a clear clinical history and histology reports. However, FFPE samples carry several limitations and considerations to be taken into account before their reliable use in microbiome research. Investigations in the quality of DNA from human FFPE samples have revealed that factors in the processing and storage (e.g. length of exposure to formalin, pH of formalin and sample storage time) negatively impact the integrity of nucleic acids and the efficacy of their downstream analyses [29, 30]. Relevant to microbiome research, unique factors to consider in quality control of FFPE samples are:

(1) Low biomass renders samples extremely susceptible to the high burden of contaminants to which they are exposed during the non-sterile FFPE processing [31]. Additionally, it aggravates the influence of host DNA, rendering samples ineffective for whole genome sequencing (WGS) and introducing PCR bias to 16S rRNA gene sequencing [32].

(2) FFPE causes DNA damage, in the form of crosslinks, DNA fragmentation, and sequence alterations [33]. In this context, 16S rRNA gene sequencing (V3-V4) necessitates DNA fragments with a length of 460 bp [34] and sequence alterations may lead to false speciation events.

(3) No sample-prep methods available for microbiome study of this sample type. FFPE microbiome studies to date have utilised approaches designed for FFPE human samples, which are suboptimal for this aim [35].

In order for the potential value of increased usage of FFPE samples for metataxonomics/metagenomics to be realised, the necessary workflows, protocols and quality control standards need to be in place [25, 36-38]. Among these, the development and systematic use of Biological Standards have been recently highlighted as a key priority for microbiome research [39-42]. Given the multiple variables (FFPE processing, storage and DNA isolation process) that directly influence the quantity and quality of DNA recovered from FFPE samples, more than perhaps any other sample type, FFPE tissue urgently requires the development of standards to ensure the validity and reproducibility of results.

A model that serves as a standard for metataxonomic and metagenomic analysis of FFPE samples requires: *1)* A defined bacterial and host cell load, *2)* Exposure to the same treatment as FFPE specimens (fixation & dehydration), *3)* A format that resembles FFPE blocks – enabling the same treatment as the source material (sectioning, deparaffination). Here is presented the Protoblock, to serve as a biological standard for FFPE samples. The Protoblock is a cell matrix, which can be populated with cell types and numbers as desired, such as to resemble those of the FFPE tissue specimens. It can be integrated in the workflow at either the FFPE processing stage for prospective studies, or at the sample prep stage for retrospective studies, allowing the assessment of either workflows, highlighting caveats that must be considered when analysing the sequencing results.

This study describes: **1)** the procedures to make the Protoblock and its validation by microscopy, **2)** validation of their value as a standard for the 16S rRNA gene sequencing and shotgun metagenomics. The Protoblock was found to be effective in enabling: ***(i)*** Assessment of currently used methods for their lysing capabilities; ***(ii)*** Measurement of the influence of host DNA and the effectiveness of host depletion strategies; ***(iii)*** Characterisation of bacterial FFPE DNA damage; ***(iv)*** Identification of common contaminants in FFPE samples

METHODS

1. Preparation of Protoblocks

Moulds. Moulds used to make a cylinder-shaped disks were made from a 54 x 11 mm adapter tube with a flat base (SARSTEDT, Cat No. 55.1570).

Cell culture. *Mus musculus* mammary gland cancer cells (4T1) were grown at 37 ^{o}C 5% CO_2, in RPMI-1640 (Sigma-Aldrich) media supplemented with 10% FBS (Sigma-Aldrich), 100 U/mL penicillin and 100 µg/mL of streptomycin (ThermoFisher), and counted with a NucleoCounter® NC-100™ (chemometect, Copenhagen).

Bacterial growth conditions. *E. coli* K12 MG1655 or *E. coli* Nissle 1917 carrying a P16Lux plasmid [43], were grown aerobically at 37 ^{o}C in Luria-Bertani (LB) medium with 300 µg/ml Erythromycin (Sigma-Aldrich). *Staphylococcus aureus* Newman (ATCC 25904) was grown aerobically at 37 ^{o}C in Todd-Hewitt broth (Sigma-Aldrich). *Bifidobacterium longum 35624* was grown anaerobically at 37 ^{o}C for 24 h in MRS medium (Sigma-Aldrich). *Lactobacillus amylophilus* (ATCC® 49845™) was grown in MRS medium (Sigma-Aldrich) at 30 ^{o}C in 5 % CO2 for 24 h. *Bacteroides thetaiotaomicron* (ATCC®29741™) was grown anaerobically at 37 ^{o}C for 24 h in FAB medium (NEOGEN, Lancashire, UK). Bacterial cultures were harvested by centrifugation and suspended in PBS. A 1 ml aliquot of the suspension was used for to count colony forming units (CFU) by retrospective plating. The rest was resuspended in Neutral Buffered Formalin and left to fix for 18 h at RT.

Counting fixed bacterial cells. The cell suspension was counted using a bacterial counting kit for flow cytometry (Invitrogen). In brief, a 10 % aliquot from the bacterial suspension was serially diluted to 1x 10^6 cells in 989 µl of NaCl. Bacterial cells were stained with 1 µl of SytoBC and 10 µl (1X 10^6) of counting beads were added to the suspension. Cells were counted in an LSR II Flow Cytometer (BD Biosciences). The acquisition trigger was set to side scatter and regulated for each bacterial strain to filter out electronic noise without missing bacterial cells. This value

was approximately 800. The volume corresponding to approximately 2 x 10^7 CFU of each bacterial strain and 2.2 x 10^7 4T1 cells were mixed together.

Fixing cells in an agar matrix. An equal volume (270 µl) of sterile agar (1.5X of elution specified by the manufacturers) pre-aliquoted and kept at 56 ºC, was pipetted into the cell suspension and thoroughly mixed by vortexing. The mixture was pipetted into the moulds, and left to solidify for 3 min at RT. Once solidified, the disk was placed in 5 ml of formalin for an extra 24 h for 48 h fixation blocks or immediately processed for 24 h fixation blocks.

Dehydration and paraffin embedding of cell disk. Cell disks were placed into a processing cassette and processed automatically with a LOGOS J (Milestone Medical, Bergamo). Here, they were dehydrated for 4 h with increasing concentrations of ethanol (37ºC), cleared 2X with xylene for 2 h 20 and 2X with isopropanol for 1 h 40 min at 37 ºC, and 1X with isopropanol for 50 min at 60 ºC. Finally, the blocks were embedded in paraffin for 8 h 32 min at 62 ºC. Once paraffinised, the Protoblocks' volume, diameter and height were measured with a calliper and by volume displacement [44]. Processed Protoblocks were placed in a 1.5 x 1.5 cm embedding mould and mounted to a processing cassette.

2. Confirmation of cell content by microscopy

Sectioning. The blocks were sectioned using aseptic technique, either at 4 µm for imaging or at 10-20 µm for DNA purification. The cell load of each slide was calculated by multiplying the total bacterial load by the volume of each slide.

Immunofluorescence and histochemistry. Cell integrity was evaluated with Gram staining (Sigma-Aldrich) or H&E staining with Mayer's haematoxylin (Sigma-Aldrich). Bacterial counts were confirmed in 3 sections stained with either 1:50 α-*E. coli* (Abcam, 137967) or 1:400 α-*S. aureus* (Abcam, 20920), counterstained with either Alexa Fluor 488 (Jackson Immunoresearch Laboratories Inc., USA) or Alexa Fluor 555-conjugated (Abcam 150062) donkey anti-rabbit Ig. Stained sections

were mounted in ProLong Gold antifade reagent with DAPI (Invitrogen, UK). Gram-stained sections were counted in bright field using an Olympus BX51 microscope, with a 100X lens. Immunofluorescent stained slides were counted at 20X (4T1 cells) or 60X (bacteria) with a fluorescence microscope (Evos FL Auto). For each slide, at least 20 randomly selected fields of view were counted. The area of the field of view (FOV) was recorded using the microscope's software and used to calculate the volume counted.

3. DNA Analysis

DNA Purification. For purifying DNA from Protoblocks, unless specified, 10 x 15 µm sections aseptically collected sections were deparaffinated with 2X xylene washes and processed following procedures specified in the QIAGEN FFPE DNA kit protocol (Qiagen Inc., Valencia, CA, USA). DNA was eluted in Tris-HCL buffer and quantified with a Qubit™ dsDNA HS Assay Kit (Invitrogen, USA). For non-fixed bacteria, bacterial cultures were grown to an OD_{600} of 1. 2 ml aliquots were processed following procedures of the GenElute™ Bacterial Genomic DNA Kit Protocol with Lysozyme and Lysostaphin (Sigma) and eluted in 50 µl of Tris-HCl. In all cases, DNA was stored at -20^{o}C until further analysis.

Fragment analysis. 1 µl of DNA purified from FFPE blocks was analysed in an Agilent 21000 bioanalyser using a High Sensitivity DNA kit (Agilent, Cat. No. 5067-4626). For Genomic Quality Number (GQN), the threshold was set to 10,000 bp and the ratio of DNA above this threshold measured for each sample. Average fragment lengths and %CV are from area underneath a maximum peak were also measured.

Quantitative PCR (qPCR). For dye-based qPCR, reactions were prepared using LUNA Universal qPCR master mix (NEB, USA) and 0.25 µM of each primer (sTable 2). Multiplex qPCR reactions were prepared using LUNA Universal Probe qPCR master mix (NEB, USA) and 0.5 µM of each primer (sTable 2) and 0.25 µM of probe for each strain. Reactions for simultaneously quantifying three bacterial strains were set using the fluorochromes: FAM, HEX, and CY3. The thermal profile included a 1 min at 95^{o}C initial denaturation, followed by 40 cycles of denaturation at 95^{o}C x 10

sec, annealing for 15 sec at the temperature specified by NEB's Ta calculator for Hot Start Taq, followed by 20-40 sec of extension at 68 ºC. For each assay, a 5-point standard curve was made from log_{10} dilutions of a gene block corresponding to species-specific genetic regions, using an initial concentration of 10^6 copies. Primers and gene-blocks were acquired from IDT (Coralville, USA) (see sTable 2 and sMaterial 1). Efficiency between 95% - 105% and R-square values > 0.995 were deemed as acceptable. All samples were run in triplicate.

qPCR Melt Curve Analysis. For melt curve analysis, FFPE *E. coli* DNA was normalised to 1 x 10^6 copies/µl. Reactions were prepared using 1X NEB Luna probe qPCR mix, 1.25 µM EvaGreen Dye (Biotium, CA, USA), 37.5 nM ROX as a reference dye, 0.25 µM of each primer (sTable 2) and 2.5 µl of template DNA. Cycling conditions used are as described for absolute quantitation with addition of a final extension step of 2 min at 68 ºC. This was followed by high-resolution melt analysis set to read fluorescence every 0.2 ºC with 10 sec soak time from 65-95 ºC. Values for the first derivative of the normalized fluorescence multiplied by -1 were exported and analysed in R environment, v3.4.4.

16S rRNA sequencing Library Preparation. Amplification of the hypervariable V3-V4 region of the 16S rRNA gene (see sTable 2) was performed in 50 µl reactions, containing 1X NEBNext High Fidelity 2X PCR Master Mix (NEB, USA), 0.5 µM of each primer, 8 µl template (5-15 ng/µl) and 12 µl nuclease free water. The thermal profile included an initial 98 ºC x 30 sec denaturation, followed by 25 cycles of denaturation at 98 ºC x 10 sec, annealing at 55 ºC x 30 sec and extension at 72 ºC x 30 sec and a final extension at 72 ºC x 5 min. Amplification was confirmed by running 5 µl of PCR product on a 2 % agarose gel. Hereafter, procedures were performed as per the Illumina 16S Metagenomic Sequencing Protocol (Illumina, CA, USA). PCR products were cleaned and sequencing libraries were prepared using the Nextera XT Index Kit (Illumina). Libraries were cleaned and quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay. Further processing was performed by GENEWIZ (Leipzig, Germany) where samples underwent a 300 bp paired-end run on the Illumina MiSeq platform.

Negative Controls. *(i)* Processing control: sterile agar exposed to the complete FFPE processing workflow. *(ii)* Wax control: wax taken from edges of an FFPE block. *(iii)* Sample prep-control was included by running an empty sample-prep reaction. *(iv)* PCR control: a 16S PCR reaction loaded with microbial DNA free water.

WGS sequencing library preparation. For NF controls, DNA from bacterial cultures of *Escherichia coli* MG1655 and *S. aureus* Newman were grown as per section 1 to and OD_{600} of 1 and their genomic DNA purified using the GenElute™ Bacterial Genomic DNA Kit Protocol with Lysozyme and Lysostaphin (Sigma). For FFPE bacteria, DNA from Protoblocks containing either strain was purified using the QIAGEN FFPE kit. In all cases DNA was eluted in 50 µl of Tris-HCl. Total purified DNA was sent to GENEWIZ (Leipzig, Germany) where WGS was performed using 2 x 150 bp chemistry on an Illlumina HiSeq.

4. Murine models

Animals, mammalian cell culture, and tumour induction. Murine experiments were approved by the Health Products Regulatory Authority (Dublin, Ireland) and the Animal Experimentation Ethics Committee of University College Cork (Cork, Ireland). RENCA cells were grown in RPMI media (Sigma) + 10% FBS (Sigma) and counted with a NucleoCounter (Chemometec). Tumours were induced in 8 week-old BALB/c mice by subcutaneous injection of 1 x 10^6 cells suspended in 200 µl serum-free RPMI media. Tumours were measured daily with a Vernier calliper and their volume calculated by measuring their longest diameter, and at the diameter perpendicular to this.

Bacterial preparation and administration. Bacteria were prepared for administration once murine tumours were approximately 5 x 5 mm in diameter. *E. coli* Nissle 1917 was grown to an OD_{600} of 0.8 in LB media, with 300 µg/ml erythromycin, harvested by centrifugation and washed 3X with PBS. *Bifidobacterium breve* UCC2003 was grown anaerobically for 24 h in Man, Rogosa, and Sharpe (MRS) media (Oxoid), + 0.05% L-cysteine hydrochloride (Sigma), harvested and washed 3X with PBS + 0.05% L-cysteine. Both bacterial strains were serially diluted to 1x10^7

CFU/ml. Tumour-bearing mice were administered 100 µl of either bacterial suspension or PBS (negative control) via lateral tail vein injection, as per [43]. Bacterial counts were confirmed by retrospectively plating in LB agar supplemented with 300 µg/ml erythromycin (*E. coli*) or RCA supplemented with 50 mg/L mupirocin (*B. breve*).

Bacterial recovery from mice. Mice were culled 7 - 11 days after bacterial administration. Tumours were aseptically excised, and halved. One half was placed in 10 % buffered formalin and fixed for 24 h at RT. The other half was placed in 1 ml PBS (+ 0.05% L-cysteine for *B. breve*) and homogenised using a 70 µm nylon cell strainer (Corning). Cell strainers were washed with 1 ml PBS. Homogenised tumours were serially diluted with PBS and plated for retrospective counting as per [45].

Formalin-fixed tissue processing. Formalin-fixed murine tissues were placed between two biopsy pads (Kaltek) in a histology cassette and processed using a LOGOS J Hybrid Tissue Processor (Milestone) and paraffin embedded as per section 1.

DNA extraction and analysis of FFPE tissue. 8 x 10 µm sections were processed for each specimen. Samples were subsequently processed with a QIAamp DNA FFPE Tissue kit (Qiagen) per the standard protocol, with the following exceptions: Tissue was deparaffinated with 2X xylene washes and the incubation with Buffer ATL and Proteinase K was performed for 1 h 45 min. DNA was eluted in 35 µl Buffer ATE. Quantitative PCR reactions were set up as per section 3, using primers and probes specified in sTable 2.

5. Bioinformatics and Statistical Analysis

Statistical analysis. All statistical analysis were performed in the R environment, v3.4.4, using methods stated in the figure legends.

16S rRNA Gene Sequence analysis. The quality of the paired-end sequence data was initially visualised using FastQC v0.11.6, and then filtered and trimmed using

Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs) and their distribution among the samples. As the aim was to characterise if contamination is present, rather than to remove it, negative controls were included to compare with the FFPE Protoblocks, with no further action taken.

Variant Calling from Whole Genome Sequence data

Filtering: HiSeq sequence data was quality filtered. Only very high quality bases were considered, to minimise the risk of sequencing errors causing false positive variants. Short fragments were also removed to reduce the likelihood of spurious alignments of regions from contaminant bacterial genomes. Trimmomatic was used to remove all reads shorter than 50 bp in length, and to trim reads when the average per base quality in a sliding window of size 4 dropped below 30.

Alignment: Of the three possible Burrows-Wheeler alignment tools, the BWA-mem aligner was used as the average read length was 150 bp, and BWA-mem is recommended when reads are over 70 bp in length as per the manual reference pages[46]. Default settings were used with the exception of allowing alignments with a minimum score of 0, rather than the default 30 as we were unsure of the extent of DNA damage induced sequence alterations. Given the stringent parameters used for read length and quality filtering, relaxing the minimum alignment score gave the best possible chance of variant detection. Samples were aligned to the original reference genome, *E. coli* MG1655.

RESULTS

1. Protoblock generation and validation.

Making the Protoblock: The Protoblock is generated by embedding a known number of fixed cells in an agar matrix that is poured into a mould that renders a defined uniform shape, in this example, a disk. Once the agar solidifies, the blocks are processed as per routine FFPE processing protocols for dehydrating and paraffin embedding, and verified by microscopy. See Figure 1.

To achieve the desired cell numbers, formalin fixed cell suspensions were counted (Figure 2, Table 1 (column 2)) and the volume of the cell suspensions normalised to cell contents (Table 1 (column 3)). For bacteria, the viable cell fraction was obtained by retrospective plating, and for murine tumour cells, viability was obtained with a NucleoCounter (Figure 1, Table 1 (Column 4)). The Protoblock radius, height, and volume were measured after dehydration. Average measurements for Protoblocks presented here were 4.99 ± 0.15 mm, 3.57 ± 0.24 mm, and 245.2 ± 14.2 µl, respectively. A slide's estimated cell population was calculated by multiplying the cell content per microliter of block (Table 1(column 5 ($\bar{x}$) & 6 (σ)) by the volume of a 15 µm slide ($\bar{x}$ = 1.57 µl, σ = 0.098 µl) or 4 µm slide ($\bar{x}$ = 0.39 µl, σ = 0.02 µl). (See Figure 1B). A slide's estimated cell population (Table 1(column 10)) was calculated by multiplying this value by the volume of a 20 µm slide ($\bar{x}$ = 1.57 µl) or 4 µm slide ($\bar{x}$ = 0.3 µl). The cell content was confirmed by immunofluorescence microscopy in blocks containing individual cell types and mixed cell content (Table 1(column 8 ($\bar{x}$) & 9 (σ)). Cell wall/membrane integrity was assessed by Gram or Haematoxylin & Eosin (H&E) staining. See Figure 2.

Protoblock validation: Protoblocks were populated with cell types and cell loads that provided the best resolution for each experimental aim. Comparable ratios of a mix of 5 bacterial strains and 4T1 cells (in the same order of magnitude ≅ 1 x 10^7) were aimed for. Estimated cell content was confirmed by immunofluorescence microscopy in blocks containing individual cell types (Figure 2C) and mixed cell content (sFigure 1). Cell wall/membrane integrity was assessed by Gram (Bacteria) or Haematoxylin

& Eosin (H&E) staining (4T1 cells). See Figure 2B. The calculated and confirmed contents for each Protoblock are specified in Figure 3 and sTable1.

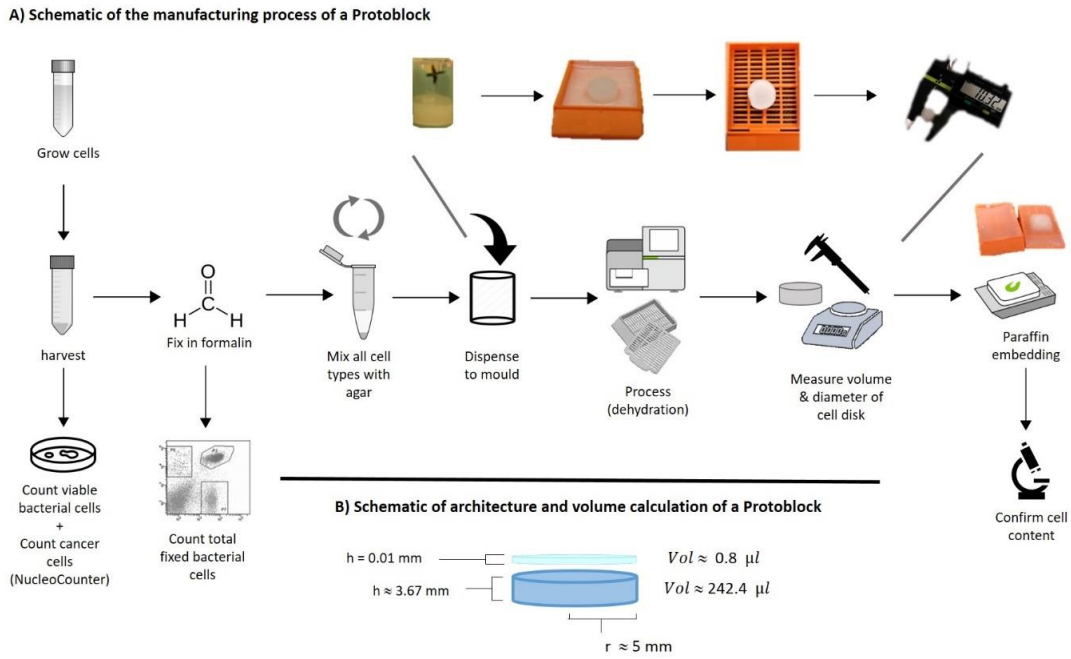


Figure 1. Making a Protoblock.

A) *Schematic of the workflow for making a Protoblock described in methods.*

B) *Schematic of the architecture of a Protoblock, demonstrating average measurements of volume, height and radius.*

Table 1. Cell input to Protoblocks and confirmation of cell contents

Column	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Cell type	*Cell susp. conc. (cells/µl)*	*Input vol. (µl)*	*% viable*	*Cell input (FACS)*	*SD Cell input (FACS)*	*Cells/µl of block*	*Cell count in block (Microsc)*	*SD Cell count in block (Microsc)*	*Cells in 15 µm slide*
4T1	*1.22E+05*	*180*	*79*	*2.20E+07*	*--*	*8.00E+04*	*2.20E+07*	*1.24E+05*	*9.07E+05*
Escherichia	*3.87E+06*	*13*	*84*	*3.94E+07*	*4.12E+06*	*1.79E+05*	*3.88E+07*	*1.32E+05*	*1.78E+06*
Staphylococcus	*1.16E+07*	*12*	*70*	*1.16E+08*	*1.55E+07*	*3.34E+05*	*9.11E+07*	*9.68E+05*	*5.04E+06*
Bifidobacterium	*3.57E+06*	*25*	*98*	*1.07E+08*	*2.38E+06*	*1.01E+05*	*9.03E+07*	*8.76E+05*	*5.02E+06*
Lactobacillus	*1.18E+07*	*11*	*100*	*1.75E+08*	*1.04E+07*	*2.27E+05*	*1.31E+08*	*1.23E+06*	*8.13E+06*
Bacteroides	*2.31E+06*	*31*	*5*	*1.23E+08*	*1.35E+06*	*4.99E+05*	*1.02E+08*	*1.42E+06*	*5.74E+06*

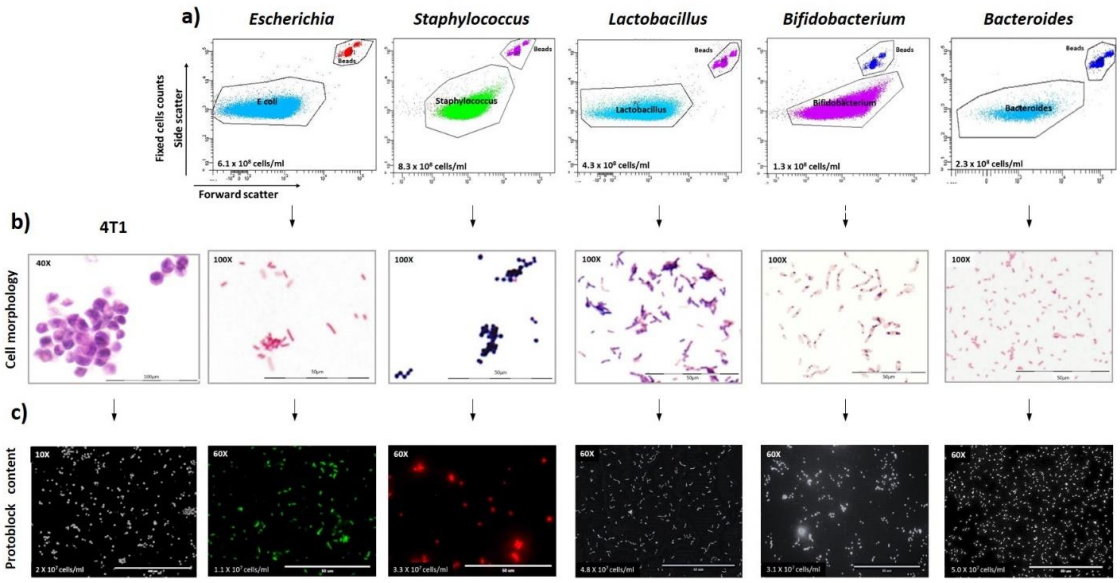


Figure 2. Validation of cell architecture and numbers in a protoblock.

a) Flow cytometry dot plots *measuring the cell density of fixed bacterial suspensions used to make protoblocks. Events were gated either for SYTOBC+ cells or beads. The averages of 3 reads for 4 populations per cell type are shown here and in Table 1.*

b) Light microscopy images *confirming cell architecture of protoblocks slides stained with H&E (4T1 cells) or Gram-staining (Bacteria).*

c) Fluorescence microscopy images *confirming cell content of protoblocks. Slides were with α- E. coli (Green), α- S. aureus (Red), or DAPI (Blue). Counts in Figure 2 are the average of 20 FOV in 3 x 4 µm slides.*

2. Protoblock for assessing bias introduced by sample prep methods

Total DNA from 10 x 15 µm slides was purified using the 'gold standard' DNA purification method for FFPE samples used in previous FFPE 16S rRNA gene sequencing studies (QIAGEN FFPE DNA kit). Protoblocks used were fixed in formalin for 24 or 48 h. Recovery by quantitative PCR was determined by quantifying the amplification of strain-specific ≅ 460bp DNA fragments (length relevant for 16S rRNA sequencing). As seen in Figure 3A (i), FFPE treated samples had at least a 10-fold reduction of amplifiable DNA, shown to be statistically significant ($p < 0.001$). Although similar amounts of DNA were purified from the samples (Figure 3C), the PCR readability of DNA is reduced by FFPE treatment, which is aggravated with increasing fixation time. Furthermore, after compensating for the 2-log fold loss of readable DNA, statistically significant under- and over-representation of all 5 genera

present was evident, with a clear bias towards Gram-negative (G-) bacteria (*Bacteroides* and *Escherichia*). This was more evident for *Bacteroides* and *Staphylococcus*, which were over- and under-represented by 605 % and - 93.1 % respectively (Figure 3A.*ii*). This effect was exacerbated by longer fixation periods. Lysis bias was confirmed with 16S rRNA gene sequencing (Figure 3 B). Altogether, these data indicate that a bacterial lysis mechanism must be incorporated in the workflow for processing of FFPE samples (this is not included in the QIAGEN kit, optimised for human DNA purification) and that for bacterial FFPE DNA, the baseline recovery of 460 bp fragments is ≤ 2-log the input. The results from these tests in Protoblocks were corroborated by FFPE murine tumour models as shown in sFigure 2.

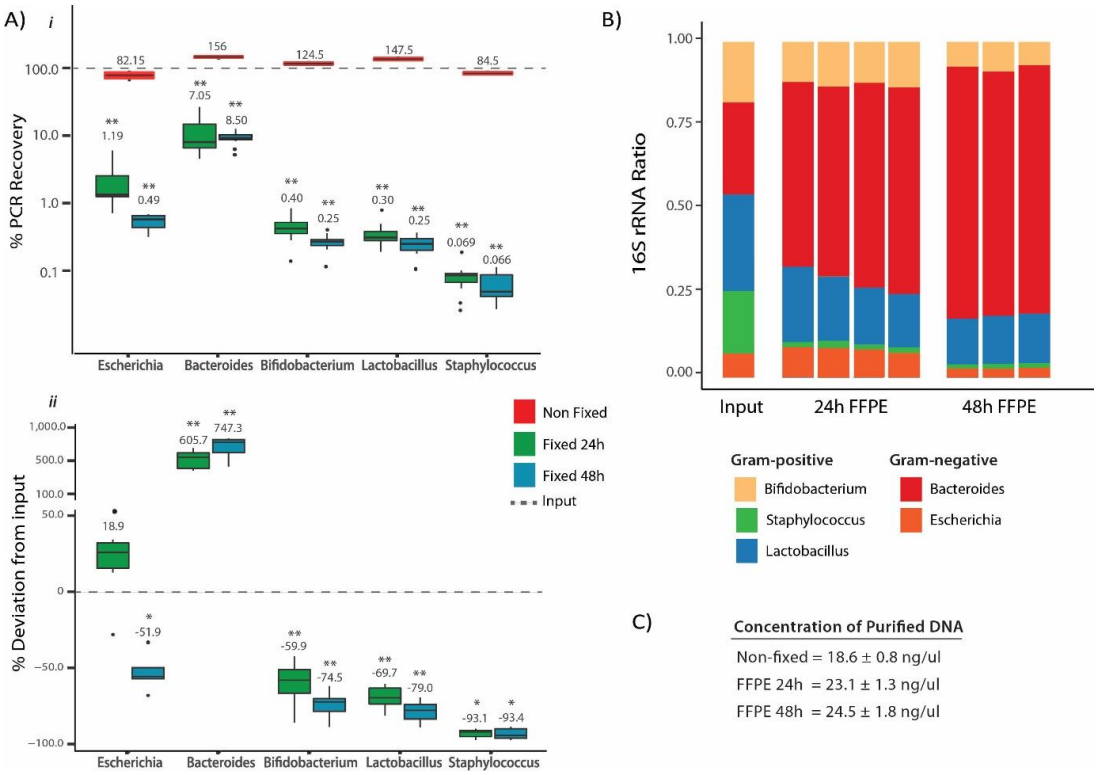


Figure 3. Assessing the recovery of FFPE bacterial DNA by quantitative PCR and 16S rRNA sequencing.

A) *Evaluating PCR recovery of FFPE bacterial DNA from Protoblocks fixed for 24 h (green) or 48 h (cyan) and compared with the recovery of paired NF samples (red).* ***i)*** *% of absolute PCR recovery (% shown above corresponding box). A 2-log fold decrease in recovery is observed for FFPE treated samples, which was found to be statistically significant in all cases as per 1 sample Wilcoxon Signed Rank test. In addition, longer fixation periods lead to a significantly greater reduction in recovery (p = 0.04).* ***ii.*** *% deviation in recovery after*

*compensating for 10-fold loss in recovery. Input = 0 (dotted line). % deviation shown above corresponding box. Significant deviation from input values, even after compensation for 10-fold decrease shown in all FFPE treated samples. (In all cases p = <0.1, * < 0.05, ** < 0.01, *** < 0.001)*

B. *Sample composition Bar plot of: Calculated input of bacterial cells added to Protoblock and 16S rRNA gene sequence analysis of Protoblocks fixed for 24 h or 48 h.*

C. *Average concentration of DNA purified from samples.*

3. Assessment of bacterial DNA integrity following FFPE

DNA fragmentation: DNA integrity was investigated with a fragment analyser by comparing DNA purified from matched NF and Protoblocks (FFPE) samples containing either a mix of NF bacteria (ratios as Table 2) or *Escherichia* only. As seen in Figure 4A, DNA fragments from NF *Escherichia* ($\bar{x}$ = 27,102 bp, %CV =65.84) or the bacterial mix ($\bar{x}$ = 31,100 bp, %CV =59.19) were highly integral (no fragmentation), with a Genomic Quality Number (GQN) > 6.6, and no significant difference was observed between sample type. On the other hand, DNA fragments from Protoblocks loaded with *Escherichia* ($\bar{x}$ = 143 bp, %CV = 41.93) or a bacteria mix ($\bar{x}$ = 110 bp, %CV =53.62) were highly fragmented with a GQN = 0.1 in both sample types. These results were in agreement with FFPE tissue DNA (sFigure 2). These results are comparable with those found in human FFPE samples, where GQN between 0.75 - 2.5 are considered high quality FFPE DNA and GQN ≤ 0.3 are low and not recommended for sequencing [47].

Assessment of PCR readable bacterial FFPE DNA: Since DNA fragmentation of FFPE bacteria was observed to be equal across taxa investigated here (Figure 4A), the effect of fragmentation on PCR recovery was investigated with Protoblocks loaded with 10^8, 10^6 and 10^4 *Escherichia* cells, as confirmed with Gram staining (Figure 4B (iii)). Quantitative PCR reactions loaded with 10^7 (61.2 +/- 5.2 ng), 10^5 (0.8 +/- 0.21 ng) or 10^3 (~ 0.02 ng) bacterial cells, were tested for the recovery of a 200 bp (recommended for FFPE) [48, 49] or 460 bp DNA fragment (required for V3-V4 16S rRNA sequencing [34]). This was compared with the recovery of a 460 bp fragment from paired NF (Non-fixed) samples (Figure 4B (i)). While comparable DNA quantities of paired FFPE/NF samples were loaded into the PCR reactions, a significant (> 1-log) reduction was observed in the quantity of DNA recovered from

Protoblock samples ($p<0.001$). A further decline in recovery (3-8 X) was evident when targeting longer (460bp) DNA fragments (Figure 4B (ii)), a trend that held true across all groups, which varied in terms of quantity of bacteria loaded, thus indicating that DNA fragmentation has a significant effect in the PCR recovery of bacterial DNA ($p<0.001$).

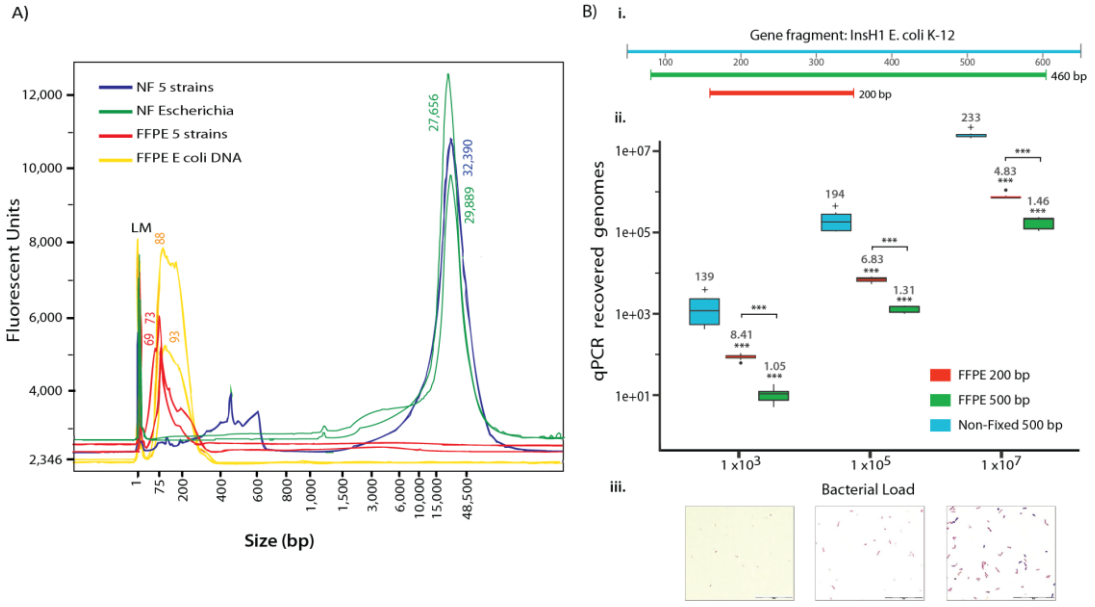


Figure 4. DNA fragmentation in FFPE bacteria

A) Evaluation of DNA integrity with fragment analyser. *Electropherograms of DNA purified from Protoblocks with a mix of 5 bacterial strains (red) and Protoblocks loaded with Escherichia only (yellow) and compared with matched NF bacterial mix (Blue) and Escherichia (Green). NF bacterial DNA had a higher integrity (GQN >6.6), while FFPE bacterial DNA from either sample was highly fragmented (GQN ≤ 0.1). No significant difference was observed between Protoblocks or NF samples. GQN = % of DNA above the threshold. The GQN threshold (dotted line) was set to that used for sequencing libraries (10,000).*

B) Measuring the recovery of PCR readable DNA from FFPE bacteria in Protoblocks by qPCR. (i) *Schematic of primer design for targeted fragments. Both 200 bp and 460 bp DNA fragments target the same E. coli K-12 regions.* ***(ii)*** *PCR recovery. Box plot of DNA recovery from 460 bp (green) and 200 bp (orange) FFPE DNA fragments (for each box, n=9) compared with NF DNA (cyan; for each box n= 6) normalised to 10^7, 10^5, 10^3 genomes. Mean recovery of DNA from Protoblocks compared with input DNA significantly differed in both FFPE sample types ($p<0.001$) as per One-Sample Wilcoxon Signed Rank test. Fragment length also significantly influenced DNA recovery of FFPE samples ($p<0.001$), as per Wilcoxon Signed Rank test.* ***(iii)*** *Gram-stained slides used for confirming bacterial content. (In all cases p = + > 0.1, . <0.1, * < 0.05, ** < 0.01, *** < 0.001)*

Presence of DNA Sequence artefacts: This was assessed in a Protoblock model populated with *E. coli*. Purified DNA was normalised to 10^6 genome copies. High resolution melt (HRM) analysis was performed in 3 contiguous DNA fragments (length ≅ 100 bp) that make up a region of the *InsH1* gene (See Figure 5A (ii)). To determine the presence of any sequence aberrations in Protoblock FFPE DNA, their melting temperature (Tm) was compared with that of NF DNA and the differences measured. Figure 5A (i) shows the final Tm for each fragment investigated. Tm shifts with variable levels of significance were observed in all fragments. Here, changes in Tm <0.1°C from that of NF DNA are indicative of low-level, non-identical sequence changes randomly distributed across the template that are typical of FFPE DNA [50]. To confirm these results, DNA purified from Protoblocks loaded with *Escherichia* and *Staphylococcus* and their paired NF samples were analysed by WGS. Findings from the DNA melting temperature analysis correlated with the results of WGS. For both bacterial strains, a higher number of sequence artefacts (chimeras and SNPs) were found in FFPE samples, when compared with their NF reference (see figure 5B).

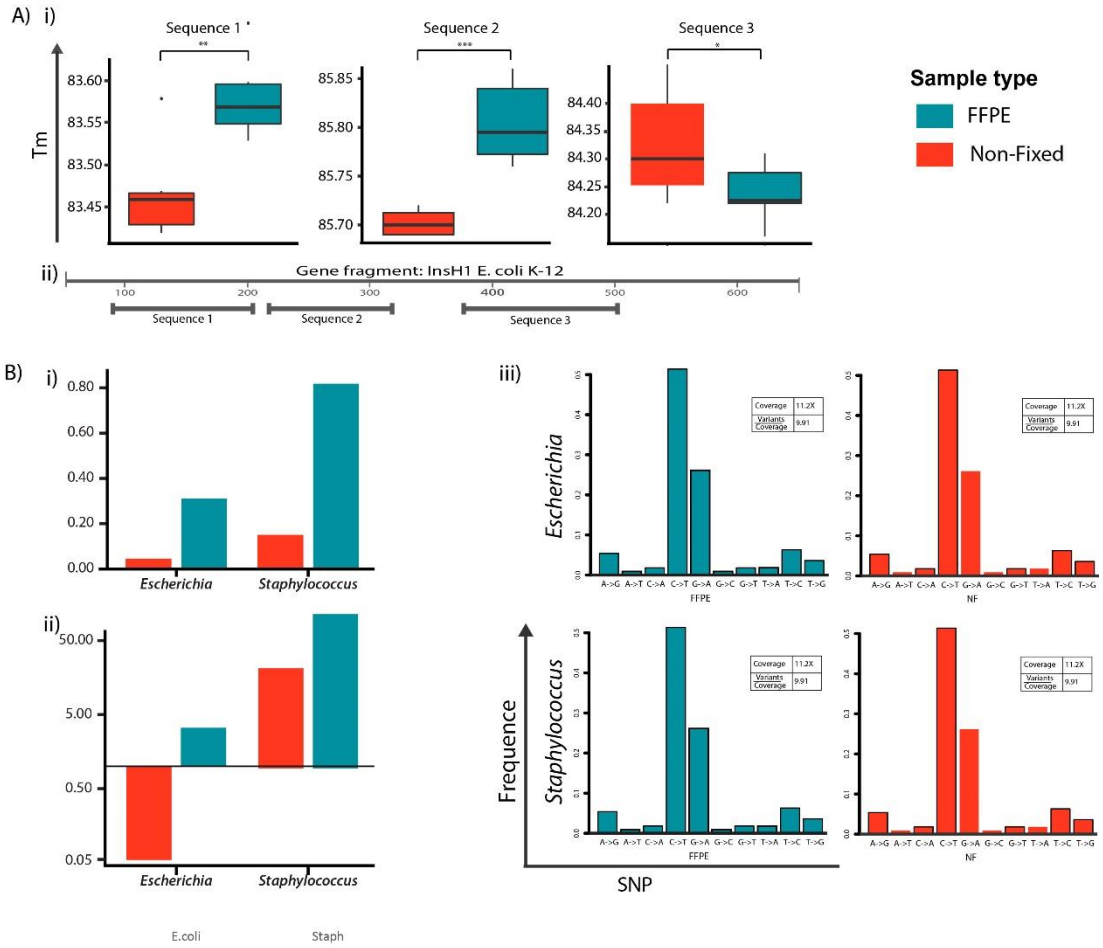


Figure 5. Evaluating sequence quality of bacterial FFPE DNA.

A) Evaluation of DNA sequence aberrations by high-resolution-melt analysis. i) *Box plots of normalised DNA quantities from Protoblock FFPE Escherichia (Cyan) and NF Escherichia (Orange). Significant shifts in the melting temperatures in 2 of the 3 sequences were observed as per Wilcoxon Signed Rank test, with temperature shifts that were on average 0.1-0.5ºC apart from NF counterparts.* ***ii)*** *Schematic of sequences used for HRM analysis: 3 DNA fragments with an average length of 100 bp were analysed. For each test and each sample type, n = 6.*

B) Confirmation of sequence alteration by WGS. *DNA from Protoblocks loaded with Escherichia and Staphylococcus and their NF paired reference was analysed by whole genome sequencing to determine chimeric reads and Single Nucleotide Polymorphisms (SNP) against the reference genome E. coli K12 MG1655 and S. aureus Newman. Here, the SNP are plotted on the x axis and the rate of occurrence on the y-axis. Variant calling, and level of coverage is measured using SAMTOOLS/BCFTOOLS. i) Chimeric reads per layer of coverage. ii) Distribution of SNPs found per bacterial strain.*

4. Characterising contaminants in the FFPE and sequencing workflow

The Protoblock is susceptible to contamination in a similar way to clinical FFPE samples. The priority of the fixing process is to preserve the tissue for later histological analysis, not to prepare a sample suitable for high throughput bacterial sequencing. In this instance, contamination was detected as shown by the number of reads in the negative controls (Figure 6). It is unlikely to have had a significant effect on the overall biological signal in this instance, given that the bacterial reads detected and their taxonomic classifications differ completely from those of the Protoblocks analysed. However, the quantity of reads detected in negative controls samples dictates that contamination remains a threat for low biomass samples, a characteristic expected in all clinically collected FFPE samples.

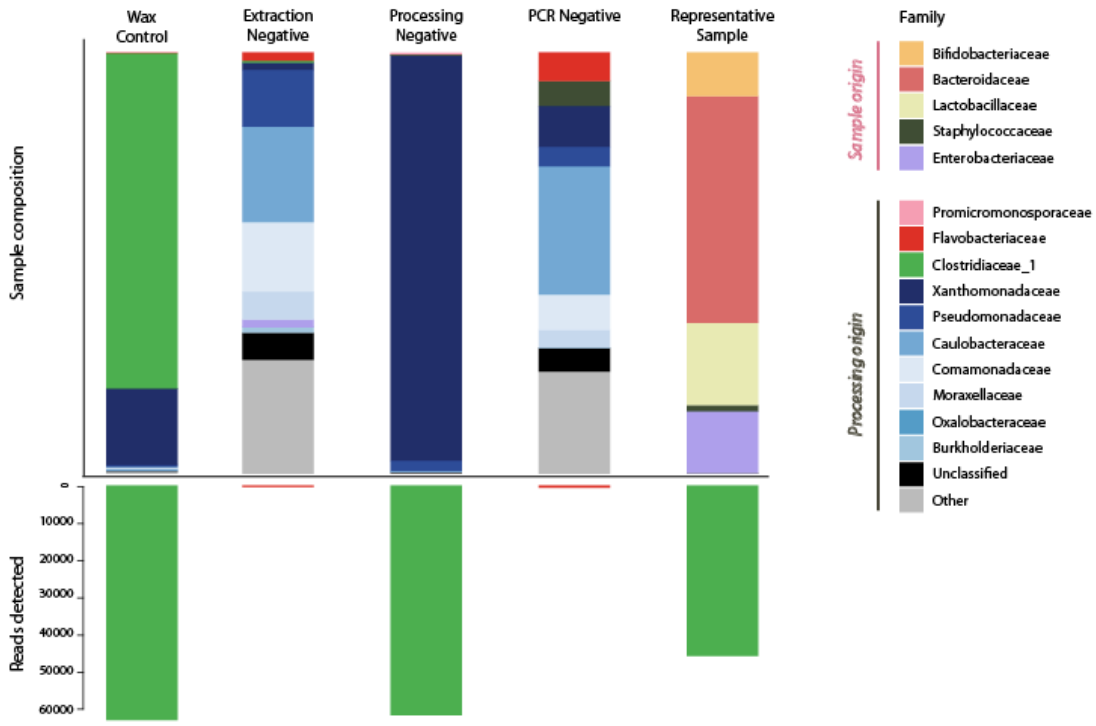


Figure 6. Evaluation of sources of environmental contamination and their effect on Protoblock samples.

Composition bar plot per sample showing proportional composition of bacterial taxa per negative control, with corresponding number of reads detected by 16S rRNA gene sequencing. Compared with representative Protoblock sample.

DISCUSSION

FFPE tissue specimens are a huge potential resource that have driven research in human cancer genomics, where numerous workflows have been developed for these samples. Over a decade of study has revealed that FFPE DNA damage is influenced by many factors during processing and storage. This results in a high inter-sample variability in the degree of DNA damage, with some samples being unsuitable for sequencing analysis [51]. To address this, the development of a robust quality control (QC) system has been crucial in directing workflows maximising the recovery, while guaranteeing the fidelity of analysis outputs. Most notable among these are the analysis of DNA fragment length (fragment analyser) and PCR readability of DNA in a sample (Infinium FFPE QC, Illumina).

Likewise, before any reliable and reproducible use of FFPE samples for microbiome analysis can be performed, a robust QC system must be developed and systematically implemented. The Protoblock presented here represents a highly relevant starting point. This method is advantageous in that the cell populations and fixation strategies can be adapted to meet the requirements for sample type and sample-prep/sequencing workflow to inform on their effects on analysis outputs [52]. Ideally such a standard would be developed in specialist facilities and distributed to researchers to guarantee sample accuracy and reproducibility across the field. This will also allow optimisation of the method to achieve a higher resemblance to tissue, such as using a larger number of host cells or incorporating extra cellular components found in tissue to the matrix. However, this method could also be adapted by researchers with specialised needs.

It has been shown here that the Protoblock is a representative FFPE model, since its contents are exposed to the same processing as FFPE experimental samples and has the same degree of DNA damage (fragmentation, PCR recovery and sequence alteration) as clinical FFPE tissue samples (Figure 4, 5 and sFigure 2). Moreover, the degree of DNA damage in the Protoblocks can be modulated by changing the severity of fixation (Figure 3A). This advantage can be exploited to develop a system similar to Infinium FFPE QC (Illumina), where a sample with a good DNA quality score serves as a standard and Cq deviations from this inform on the suitability of samples for sequencing analysis. The Protoblock can also serve as a quantitative standard to

determine cycle number at which tested FFPE samples will have detectable levels of 16S rRNA gene sequences, if any.

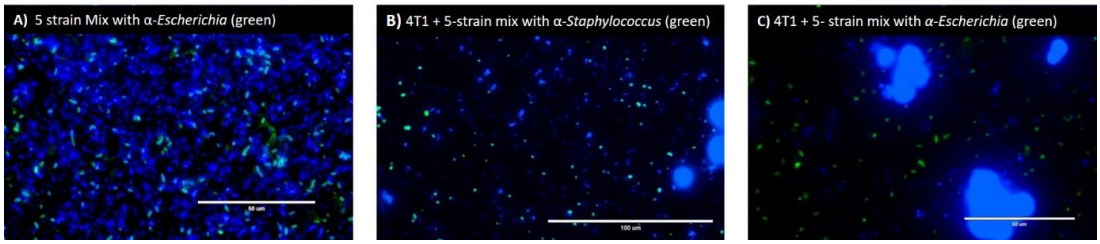
From the results shown here, it is clear the QIAGEN FFPE DNA sample prep is unsuitable for microbiome analysis, since it is strongly biased towards Gram-negative bacteria (Figure 3). Given the lack of a standardised method to process FFPE samples for metagenomic studies, the use of standards such as the Protoblock is essential to develop this workflow and guarantee the accuracy, precision, and limit of detection of the analysis. An unexpected finding was a higher than expected recovery of FFPE *Bifidobacterium* in samples processed without undergoing bacterial lysis. The opposite was found for *Staphylococcus.* This reinforces the need to thoroughly study the effect of FFPE on bacteria prior to any microbiome analysis of FFPE specimens. Principally, a thorough investigation on the effect of FFPE in bacterial membrane/cell walls and bacterial DNA itself.

Finally, contamination is a considerable threat to the accuracy of sequence-based analysis of low biomass samples such as FFPE specimens. Steps in the processing of FFPE samples require the use of solutions that are difficult to keep sterile, and contamination from these sources could easily obscure the true results in cases of low microbial load. Use of a combination of an empty (agar only) and a bacterial loaded Protoblock along with a sample of the paraffin wax used for embedding can inform on the most common contaminants and the level of contamination introduced by any processing of FFPE samples required, in advance of a sequencing study. Although, contamination was minimal, due to sufficient bacterial biomass, clinically collected FFPE blocks can be expected to have a much lower level of microbial biomass and are thus more susceptible.

CONCLUSION

Unlocking the potential of FFPE samples for microbiome analysis could have a huge effect on the field. For this to be a reality, a robust quality control system needs to be developed. The Protoblock presented in this study is foundational in building towards this. Evidence generated here shows its value in investigating the effect of FFPE in bacteria and optimisation of sequencing workflows for this sample type.

SUPPLEMENTARY MATERIAL

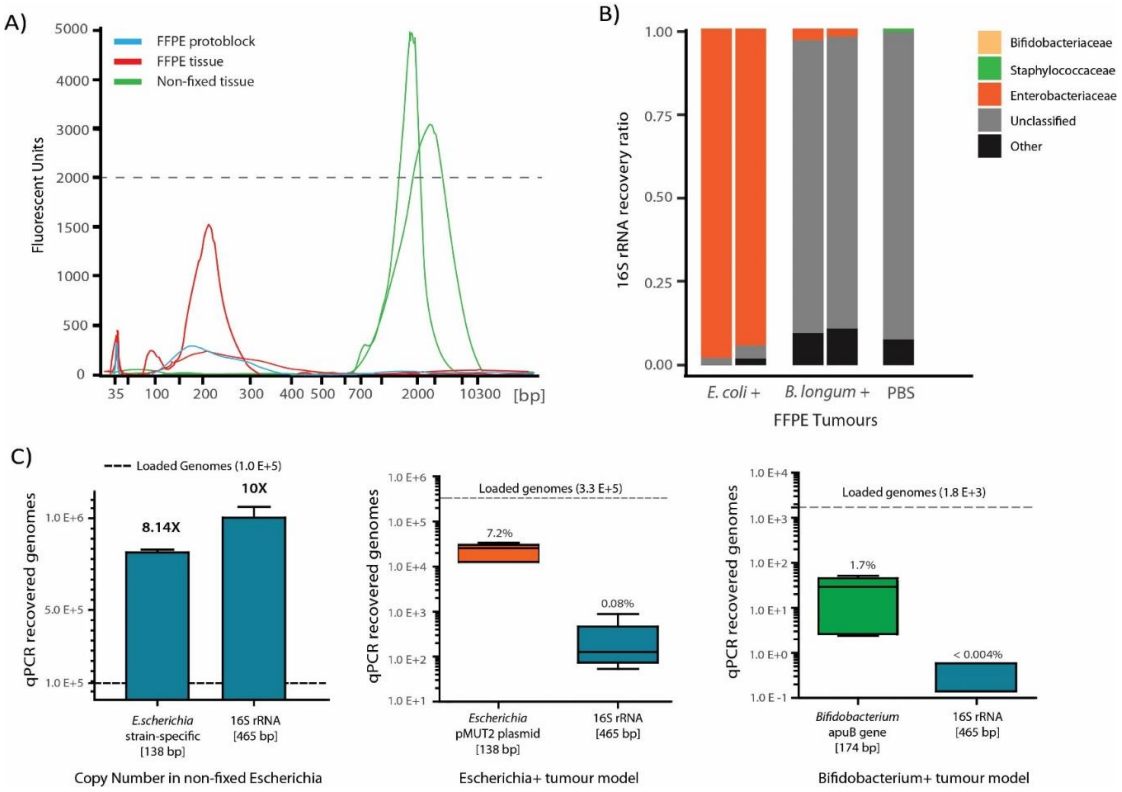


Supplementary Figure 1. Microscope images of Protoblocks with 5-strain mix contents

A) *Microscope image (40X) of Protoblock loaded with the 5 bacterial taxa specified in Figure 2. DAPI (Blue), staining all bacterial cells. In green, α-Escherichia.*

B) *Microscope image (40X) of Protoblock loaded with the 4T1 cells and the 5 bacterial taxa specified in Figure 2. DAPI (Blue), staining 4T1 and bacterial cells. In green, α-Staphylococcus.*

B) *Microscope image (40X) of Protoblock loaded with the 4T1 cells and the 5 bacterial taxa specified in Figure 2. DAPI (Blue), staining 4T1 and bacterial cells. In green, α-Escherichia.*



Supplementary Figure 2. Validation of findings in FFPE tissue.

A) DNA fragmentation. *Electropherograms comparing the integrity of NF tissue DNA (green), with FFPE tissue DNA (red) and the contents of Protoblocks (blue). NF tissue fragment length = 4,406 ± 1,939 bp, DNA from FFPE tissue = 229 ± 20 bp and DNA from Protoblocks = 192 ± 48.5 bp.*

B) 16S rRNA recovery of FFPE tissue. *Bar plot showing bacteria recovered by 16S rRNA sequencing from murine tumours models loaded with either Escherichia, Bifidobacterium or PBS. Here, Escherichia was readily detected, while Bifidobacterium was not detected.*

C) ***PCR recovery of Escherichia and Bifidobacterium from FFPE tissue. i) Assessment of strain specific gene and 16S rRNA gene in the recovery of non-fixed E. coli.*** *Bar plot showing the number of gene copies retrieved for either a strain specific gene ($\bar{x}$ = 8.14 ± 0.43 copies) or 16S rRNA gene ($\bar{x}$ = 10 ± 1.2 copies) after amplifying an input of 1 x 10^5 Escherichia cells.* ***ii) Recovery of Escherichia from FFPE tumours****. Box plot showing the PCR recovery of an input of 3.3 x 10^5 Escherichia genomes with a strain specific 137 bp DNA fragment (2.33 x 10^4 ± 8.8 x 10^4 genomes) and that of the 16SrRNA gene (2.7 x 10^2 ± 3.2 x 10^2 genomes).* ***iii) Recovery of Bifidobacterium from FFPE tumours.*** *Box plot showing the PCR recovery of an input of 1.8 x 10^3 Bifidobacterium genomes with a strain specific 174 bp DNA fragment (3.2 x 10^1 ± 21 x 10^1 genomes) and that of the 16SrRNA gene for which there was not a reliable amplification detected (only 2/6 replicates returned an average amplification of 7 x 10^{-2} copies.*

Supplementary Table 1. Primers and Probes used for Protoblock analysis

Strain/Cell line	Gene/ Accession No	Primer/Probe sequence	F/R/P	Product size (bp)	Figure
E coli MG1655 [CP032667]	IS5-like element IS5 family transposase AYG17556.1 [CP032667: 230175-231191]	5'TCA TTT GGT CCG CCC GAA AC	F	525	4B, 5B
		5'CCA CCA TCA TTG AGG CAC CC	R		
		5'GCC GAA CTG TCG CTT GAT GA	F	217	4A, 4C, 5B
		5'ATT TGT CTC AGC CGA TGC CG	R		
		5'TCG GCT GAG ACA AAT TGC TC	F	110	6A(i)
		5'GAT GCC AAG AGT GGC CTG	R		
		5'ATG CCA AAG TGC CAC TGA T	F	100	6A(ii)
		5'CCA CCA TCA TTG AGG CAC C	R		
		5'CCC CTT GTA TCT GGC TTT CA	F	116	6A(iii)
		5'AGA ACA AAA CGG CCA TCA AC	R		
E coli Nissle 1917 [CP022687]	plasmid pMUT2 [CP022687] [53]	5'GAA CAT ACA GAC CGC TAT CC	F	460	3A,5B
		5'GCC TCT GTA AGC TCT CTA ATG	R		
		56-FAM/CTTGATGAC/ZEN/CTGACGATGTTGAGC/3IABkFQ/	P		
		5'-AACACTGGAATATGTGGCCCAAAG	F	137	sF1
		5'-GGGCTCGGGGATCAAATTCAAG	R		
		5'-/56-FAM/AGCCATCAA/ZEN/ATCGGCATCATCCTCGGT/3IABkFQ/-3'	P		
Staphylococcus aureus subsp. *aureus* str. Newman [CP023390.1]	Thermonuclease ATC67584.1 [CP023390.1:1359312-1359845] [54]	5'TGC TAT GAT TGT GGT AGC CAT C	F	425	3A,4B,
		5'ACT TCT CTC TAG CAA GTC CCT	R		
		5'Cy3/CAA GAT CGC TAT GGT AGA ACA TTG GCG TAT G/3BHQ_2/	P		
		5'CGC CTG TAC AAC CAT TTG GC	F	182	4A, 4C
		TCT AGC AAGT CCC TTT TCC ACT	R		
Lactobacillus amylophilus (ATCC® 49845™) [1423721]	hsp60 gene [HE573891.1: 314172 - 314752] [55]	5'CCC TTG GAA CGT GGT TAT G	F	474	3A
		5'ACG GGT TCC TTC GACT T	R		
		5HEX/CCA TTA AAC/ZEN/AGG GTC GTG GGT ATG/3IABkFQ/	P		
Bacteroides thetaiotaomicron	NP_809948.1 [NC_004663.1: 1306764-1307540] [56]	5'CAT TCT TCT TCT TGT GGC TAA AC	F	480	3A,5B
		5'TGG GAA ATG TAC AAC CTG AAA	R		

(ATCC®29741™) NZ_CP012937.1		56-FAM/TGA GCT TGG/ZEN/GCT ATT TGC TGT TTA/3IABkFQ	p		
Bifidobacterium longum strain 35624	Glycosyl transferase CP013673 [461108-461924][57]	5'CGT CGT CGT CTG ATT CGT AAG	F	440	4A
		5'GGG CGC TTG ATA GAG AAC AA	R		
		5'HEX/CTA TAA GGT /ZEN/AAA TCT TCC AGC CGT ACC GGA G/3IABkFQ/	P		
		5'GTC GGA CTT GCT GCG TTT ATC GTT G	F	125	4C
		5'CGG GGC GCT TGA TAG AGA ACA ATG	R		
4T1 cells [ATCC® CRL-2539™] *Mus musculus* [10090]	BetaActin AC144818.4 [NC000071.6: 73696- 73082]	5'GAT TAC TGC TCT GGC TCC TAG	F	147	4C
		5'GAC TCA TCG TAC TCC TGC TTG	R		
		5'/HEX/CTG GCC TCA/ZEN/CTG TCC ACC TTC C/3IABkFQ/	p		
V3-V4 hypervariable region of 16S rRNA gene	341F_Overhang and 785R_Overhang	5'TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG	F	460	3B
		5'GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CHV GGG TAT CTA ATCC	R		
Bifidobacterium breve UCC2003	*apuB* gene	5'-GCAATGGATCAAGACGTTCG	F	174	sF1
		5'-TCATACGGTGCCCAAAAGG	R		
		5'-/56-FAM/AGCAGTTGG/ZEN/CGAAGATCACCGA/3IABkFQ	P		

Troubleshooting/Technical Considerations

Key elements to be considered when creating accurate Protoblock models are a precise cell load estimation, maintenance of cell integrity and a shape that facilitates a uniform cell distribution that allows for slides to be representative of the block's populations. To achieve this, the following points must be considered when preparing blocks.

Cell counts. The protoblock consists of cell populations of viable and non-viable cells. Downstream quantification of bacterial content via microscopy and qPCR informs the total bacterial content, with no distinction between viable and non-viable cells. Moreover, DNA from non-viable cells is more readily accessible and more easily recovered during DNA isolation, which could introduce bias in the downstream analyses. To avoid this, it is important that cell counts are done on the fixed population as a whole, with the viable cell content calculated to estimate bias introduced by readily accessible DNA during sample preparation. Flow cytometry and fluorescent microscopy were deemed as feasible approaches to obtain a total bacterial count.

Cell displacement volume. Displacement volume can have a significant effect on total cell counts and must be taken into account. The volume used to re-suspend the cells will not represent the final volume of cell suspension if the displacement volume is not considered. Higher effects were observed for mammalian cells (2.4×10^{-6} µl/cell) and bacteria excreting exopolysaccharides, such as *Bifidobacterium* (1×10^{-6} µl/cell). It is therefore essential that the final volume of cell suspension is confirmed when calculating cell density. Displacement volumes for the strains used in the Protoblocks presented here are outlined in the material and methods.

Volume of the block: During dehydration the water content of the Protoblock is removed and the volume of the block significantly reduced. The cell density of the block in terms of volume and diameter, must therefore be calculated in dehydrated blocks. In addition, dehydration enhances the presence of a meniscus, which also has an effect on the final volume. To account for this, the volume should also be measured using a method designed for irregular objects. Here, the blocks were measured successfully with the Archimedes' principle, wherein their volume is equal to the volume of water they displace. Given the small volume of the block, it is important to take repeated measurements and confirm the volume measurements manually using Vernier calliper measurement.

Maintaining cell integrity: The accuracy of the Protoblocks as a standard is also determined by the cell integrity of its population. This ensures an accurate representation of the fixation and purification processes carried out. . In order to ensure cell integrity, there are 3 key aspects to consider: 1) Centrifugation must be kept to a minimum and adjusted to the lower speed settings (<5000 *x g*) to allow pelleting of each cell type. 2) Formalin fixation should be performed immediately after harvesting the cells. 3) The molten agar must be kept below 60 °C when embedding the cells

Maintaining the shape of the block: The shape of the block must be maintained during processing to ensure an accurate representation of the cell population throughout the block. This can be ensured by: 1) Using a higher concentration of agar to compensate for the input volume of cells, 2) Incubating the cells at 50°C for 1 min before mixing with the agar to prevent solidification before placing in the mould, 3) Keeping agar aliquots with 30 -50 µl more than the desired volume at 60°C to avoid evaporation,

prevent solidification of the agar during pipetting and avoid bubble formation. 4) Swirling the mould after depositing the mix to ensure an even distribution. 5) Removing any bubbles with a bacterial loop before it solidifies 6) Using filter paper to protect the block inside the cassette. 7) Embedding the block with the bottom face at the base of the block and ensure that sections correspond to full face sections. To avoid oversaturating the reactions with paraffin, use smaller (2 x 2 cm) embedding moulds.

Considerations during DNA isolation: Cells are embedded in an agar matrix. To avoid interference of the agar with column-based DNA isolation procedures, centrifuge the samples at 17,000 x g for 1 min before transferring to spin columns. Ensure to avoid contamination of the pipette with agar during this process.

REFERENCES

1. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* Journal of Experimental Biology, 2007. **210**(9): p. 1518-1525.

2. Forbes, J.D., et al., *Metagenomics: The Next Culture-Independent Game Changer.* Frontiers in Microbiology, 2017. **8**(1069).

3. Knight, R., et al., *Best practices for analysing microbiomes.* Nature Reviews Microbiology, 2018. **16**(7): p. 410-422.

4. Castillo, D.J., et al., *The Healthy Human Blood Microbiome: Fact or Fiction?* Frontiers in Cellular and Infection Microbiology, 2019. **9**(148).

5. Stinson, L.F., et al., *The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth.* Frontiers in Microbiology, 2019. **10**(1124).

6. Ozkan, J., et al., *Identification and Visualization of a Distinct Microbiome in Ocular Surface Conjunctival Tissue.* Investigative Ophthalmology & Visual Science, 2018. **59**(10): p. 4268-4276.

7. Zhou, B., et al., *The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients.* Scientific Reports, 2019. **9**(1): p. 1691.

8. Chen, J., et al., *The microbiome and breast cancer: a review.* Breast Cancer Res Treat, 2019.

9. Beck, J.M., V.B. Young, and G.B. Huffnagle, *The microbiome of the lung.* Translational research : the journal of laboratory and clinical medicine, 2012. **160**(4): p. 258-266.

10. Huffnagle, G.B., R.P. Dickson, and N.W. Lukacs, *The respiratory tract microbiome and lung inflammation: a two-way street.* Mucosal Immunology, 2017. **10**(2): p. 299-306.

11. Marsh, R.L., et al., *The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx.* Microbiome, 2016. **4**(1): p. 37.

12. Emery, D.C., et al., *16S rRNA Next Generation Sequencing Analysis Shows Bacteria in Alzheimer's Post-Mortem Brain.* Frontiers in Aging Neuroscience, 2017. **9**(195).

13. Stewart, C.J., et al., *Using formalin fixed paraffin embedded tissue to characterize the preterm gut microbiota in necrotising enterocolitis and spontaneous isolated perforation using marginal and diseased tissue.* BMC Microbiology, 2019. **19**(1): p. 52.

14. Hart, J.D., et al., *16S rRNA sequencing in molecular microbiological diagnosis of bacterial infections in the autopsy setting.* Pathology, 2014. **46**: p. S113.

15. Racsa, L.D., et al., *Identification of bacterial pathogens from formalin-fixed, paraffin-embedded tissues by using 16S sequencing: retrospective correlation of results to clinicians' responses.* Human Pathology, 2017. **59**: p. 132-138.

16. Banerjee, S., et al., *Distinct Microbial Signatures Associated With Different Breast Cancer Types.* Frontiers in Microbiology, 2018. **9**(951).

17. Baldwin, D.A., et al., *Metagenomic assay for identification of microbial pathogens in tumor tissues.* mBio, 2014. **5**(5): p. e01714.

18. Riquelme, E., et al., *Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes.* Cell, 2019. **178**(4): p. 795-806.e12.

19. Xuan, C., et al., *Microbial Dysbiosis Is Associated with Human Breast Cancer.* PLoS ONE, 2014. **9**(1): p. e83744.

20. Gaffney, E.F., et al., *Factors that drive the increasing use of FFPE tissue in basic and translational cancer research.* Biotechnic & Histochemistry, 2018. **93**(5): p. 373-386.

21. van Beers, E.H., et al., *A multiplex PCR predictor for aCGH success of FFPE samples.* British journal of cancer, 2006. **94**(2): p. 333-337.

22. Blow, N., *Tissue issues.* Nature, 2007. **448**(7156): p. 959-960.

23. Kerick, M., et al., *Targeted high throughput sequencing in clinical cancer Settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity.* BMC Medical Genomics, 2011. **4**(1): p. 68.

24. Hedegaard, J., et al., *Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue.* PloS one, 2014. **9**(5): p. e98187-e98187.

25. McDonough, S.J., et al., *Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods.* PloS one, 2019. **14**(4): p. e0211400-e0211400.

26. Kresse, S.H., et al., *Evaluation of commercial DNA and RNA extraction methods for high-throughput sequencing of FFPE samples.* PLOS ONE, 2018. **13**(5): p. e0197456.

27. Siebolts, U., et al., *Tissues from routine pathology archives are suitable for microRNA analyses by quantitative PCR.* Journal of Clinical Pathology, 2009. **62**(1): p. 84-88.

28. Fanelli, M., et al., *Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue.* Nature Protocols, 2011. **6**(12): p. 1905-1919.

29. Srinivasan, M., D. Sedmak, and S. Jewell, *Effect of fixatives and tissue processing on the content and integrity of nucleic acids.* The American journal of pathology, 2002. **161**(6): p. 1961-1971.

30. Vitosevic, K., et al., *Effect of formalin fixation on pcr amplification of DNA isolated from healthy autopsy tissues.* Acta Histochem, 2018. **120**(8): p. 780-788.

31. Hykin, S.M., K. Bi, and J.A. McGuire, *Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing.* PloS one, 2015. **10**(10): p. e0141579-e0141579.

32. Feehery, G.R., et al., *A method for selectively enriching microbial DNA from contaminating vertebrate host DNA.* PloS one, 2013. **8**(10): p. e76096-e76096.

33. Do, H. and A. Dobrovic, *Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization.* Clinical Chemistry, 2015. **61**(1): p. 64-71.

34. Bukin, Y.S., et al., *The effect of 16S rRNA region choice on bacterial community metabarcoding results.* Scientific Data, 2019. **6**(1): p. 190007.

35. Ruiz, L., et al., *How do bifidobacteria counteract environmental challenges? Mechanisms involved and physiological consequences.* Genes & nutrition, 2011. **6**(3): p. 307-318.

36. Einaga, N., et al., *Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.* PLOS ONE, 2017. **12**(5): p. e0176280.

37. Bailey, S.T., et al., *High-quality whole-genome sequencing of FFPE samples.* Journal of Clinical Oncology, 2018. **36**(15_suppl): p. e13500-e13500.

38. Bettoni, F., et al., *A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology.* Exp Mol Pathol, 2017. **103**(3): p. 294-299.

39. Costea, P.I., et al., *Enterotypes in the landscape of gut microbial community composition.* Nature Microbiology, 2018. **3**(1): p. 8-16.

40. Proctor, L., *Priorities for the next 10 years of human microbiome research.* Nature, 2019. **569**(7758): p. 623-625.

41. Costea, P.I., et al., *Towards standards for human fecal sample processing in metagenomic studies.* Nature Biotechnology, 2017. **35**: p. 1069.

42. Katsnelson, A., *Standards Seekers Put the Human Microbiome in Their Sights.* ACS Cent Sci, 2019. **5**(6): p. 929-932.

43. Cronin, M., et al., *High resolution in vivo bioluminescent imaging for the study of bacterial tumour targeting.* PLoS One, 2012. **7**(1): p. e30940.

44. Hughes, S. and J. Lau, *A technique for fast and accurate measurement of hand volumes using Archimedes' principle.* Australasian Physics & Engineering Sciences in Medicine, 2008. **31**(1): p. 56.

45. Cronin, M., et al., *Orally administered bifidobacteria as vehicles for delivery of agents to systemic tumors.* Mol Ther, 2010. **18**(7): p. 1397-407.

46. Li, H. *bwa - Burrows-Wheeler Alignment Tool*. 2013; Available from: http://bio-bwa.sourceforge.net/.

47. Illumina, I., *Evaluating DNA Quality from FFPE Samples*, in *Technical Note: Library Preparation*. 2016, Illumina, Inc.

48. Lindahl, T., *Instability and decay of the primary structure of DNA.* Nature, 1993. **362**(6422): p. 709-715.

49. Vesnaver, G., et al., *Influence of abasic and anucleosidic sites on the stability, conformation, and melting behavior of a DNA duplex: correlations of thermodynamic and structural data.* Proceedings of the National Academy of Sciences, 1989. **86**(10): p. 3614-3618.

50. Do, H. and A. Dobrovic, *Limited copy number-high resolution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies.* Mol Cancer, 2009. **8**: p. 82.

51. Robbe, P., et al., *Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.* Genet Med, 2018. **20**(10): p. 1196-1205.

52. Choo, J.M., L.E.X. Leong, and G.B. Rogers, *Sample storage conditions significantly influence faecal microbiome profiles.* Scientific Reports, 2015. **5**: p. 16350.

53. Blum-Oehler, G., et al., *Development of strain-specific PCR reactions for the detection of the probiotic Escherichia coli strain Nissle 1917 in fecal samples.* Res Microbiol, 2003. **154**(1): p. 59-66.

54. Madison, B.M. and V.S. Baselski, *Rapid identification of Staphylococcus aureus in blood cultures by thermonuclease testing.* Journal of clinical microbiology, 1983. **18**(3): p. 722-724.

55. Blaiotta, G., et al., *Lactobacillus strain diversity based on partial hsp60 gene sequences and design of PCR-restriction fragment length polymorphism assays for species identification and differentiation.* Applied and environmental microbiology, 2008. **74**(1): p. 208-215.

56. Teng, L.J., et al., *PCR assay for species-specific identification of Bacteroides thetaiotaomicron.* J Clin Microbiol, 2000. **38**(4): p. 1672-5.

57. Altmann, F., et al., *Genome Analysis and Characterisation of the Exopolysaccharide Produced by Bifidobacterium longum subsp. longum 35624.* PLoS One, 2016. **11**(9): p. e0162983.

CHAPTER 3:

Characterisation of FFPE-Induced Bacterial DNA Damage and Development of a DNA Repair Method for Metagenomics & metataxonomics

ABSTRACT

Formalin-fixed, paraffin-embedded (FFPE) samples have huge potential as source material in the field of human microbiome research. However, the effects of FFPE processing on bacterial DNA remain uncharacterised. Any effects are relevant for microbiome studies, where DNA template is often minimal and sequences studied are not limited to one genome. As such, we aimed to (i) characterise FFPE-induced bacterial DNA damage, and (ii) develop strategies to reduce and repair this damage.

Our analyses indicate that bacterial FFPE DNA is highly fragmented, a poor template for PCR, crosslinked and bears sequence artefacts derived predominantly from oxidative DNA damage. Two strategies to reduce this damage were devised - an optimised decrosslinking procedure reducing sequence artefacts generated by high-temperature incubation, and secondly, an *in vitro* reconstitution of the Base Excision Repair (BER) pathway. As evidenced by whole genome sequencing, treatment with these strategies resulted in 3X increase in fragment length and a significant reduction in sequence artefacts. This translated to an increased sequencing readability. Application of this strategy to mammalian FFPE DNA produced similar improvements.

This study provides a new understanding of the condition of bacterial DNA in FFPE specimens and how this impacts downstream analyses, in addition to a strategy to improve the sequencing quality of bacterial and mammalian FFPE DNA.

INTRODUCTION

Formalin fixed paraffin embedded (FFPE) samples represent the most comprehensive collections of patient materials in hospital pathology archives [1-3]. These samples can provide access to bacterial communities inhabiting a variety of body sites for which access to 'fresh' tissue samples is limited [4, 5] due to the invasive nature of their sampling [6-11]. However, as has been definitively shown from analysis of human DNA [12], FFPE processing induces DNA damage. In mammalian DNA, this damage occurs as: **(i)** Cross-links (DNA-DNA, Protein-DNA) [13, 14], **(ii)** Depurination [15-17], **(iii)** DNA fragmentation [18, 19], and **(iv)** Sequence alterations (chimeras, SNPs) [20, 21], which accumulate further with storage time and suboptimal fixing conditions [12, 22]. This DNA damage has been found to negatively affect mammalian DNA sequencing outputs, by reducing: *a)* the sequencing depth, *b)* sequencing uniformity, *c)* read length, *d)* ratio of reads passing quality filtering; and increasing *a)* the number of chimeric reads, *b)* FFPE derived single nucleotide polymorphisms (SNPs), translocations, and insertions and deletions (indels) [12, 23-29].

Bacterial DNA is likely to be similarly damaged, but this is uncharacterised to date. The consequence of such bacterial DNA damage is that FFPE samples will have several associated limitations that must be considered before their effective use in microbiome studies. DNA fragmentation reduces the quantity of DNA fragments within a sample of suitable length for amplicon-based sequencing strategies such as 16S rRNA gene sequencing (~460 bp for V3-V4 [30]). This can exacerbate the characteristic low bacterial biomass found in FFPE samples. FFPE-induced sequence alterations can decrease sequence quality and lead to false speciation events. These are considerable hurdles standing in the way of accurate, reproducible microbiome research from FFPE samples.

All research reported to date, and protocols for purifying and repairing FFPE DNA, relate to mammalian (human) DNA. Differences in DNA conformation and packaging, methylation patterns, and replication and transcription rates, between human and bacteria may lead to different FFPE damage profiles [31-33]. A better understanding of potential differences is essential for the proper design of workflows

that ensure bacterial DNA quality and guarantee reliable and reproducible sequencing analysis [34]. No characterisation of FFPE-induced bacterial DNA damage exists to date.

Assuming the existence of such damage, the Base Excision Repair (BER) pathway represents a promising opportunity to repair it before subjecting it to analyses. BER is the main cellular pathway for repair of lesions, such as damaged bases, AP sites and ss-breaks [35, 36]. Strategies to improve the sequencing quality of FFPE human samples using an individual enzyme from the BER pathway have been adopted - namely, Uracil DNA glycosylase [37]. In addition, commercial kits for some degree of FFPE DNA repair have recently become available: 'NEB FFPE DNA Repair' and 'Illumina Infinium FFPE Repair'; however, their composition is undisclosed. Despite such advances, there is a gap in the literature characterising DNA damage recognition by DNA glycosylases on FFPE samples, which is essential for designing approaches to reconstitute the BER pathway to repair FFPE DNA damage. To our knowledge, the only reports available were designed to assess the outcomes of human DNA repair after treatment with a commercial kit [38].

The BER pathway can be summarised in 5 steps. i) Base excision by a DNA glycosylase, followed by ii) backbone excision by an AP lyase, iii) ends processing by a polynucleotide kinase or exonuclease, iv) gap filling by a polymerase, and v) nick ligation by a ligase [35, 36]. The type of DNA glycosylase determines downstream repair workflow. Excisions made by monofunctional DNA glycosylases are repaired through long-patch BER [39, 40], and excisions made by bifunctional glycosylases, through short-patch BER [35, 36, 39-43].

In this study, a 'mock' FFPE model replicating the conditions found in clinical FFPE samples, was used to characterise the nature and severity of FFPE-induced damage in bacterial DNA, followed by development of an effective strategy for repairing it. Quantitative PCR and high resolution melt analysis, along with Sanger Sequencing were used to screen a set of available DNA repair enzymes, and shortlist those found most effective. These were then further tested individually and in combination, with a final validation of Whole Genome Sequencing (WGS) analysis used to determine the most effective DNA repair strategy.

METHODS

1. Preparation of FFPE blocks

Bacterial growth conditions. *E. coli* K12 MG1655 or *E. coli* Nissle 1917 carrying a P16Lux plasmid [44], were grown aerobically at 37 ºC in Luria-Bertani (LB) medium with 300 µg/ml Erythromycin (Sigma-Aldrich). *Staphylococcus aureus* Newman (ATCC 25904) was grown aerobically at 37 ºC in Todd-Hewitt broth (Sigma-Aldrich). *Bifidobacterium longum 35624* was grown anaerobically at 37 ºC for 24 h in MRS medium (Sigma-Aldrich). *Lactobacillus amylophilus* (ATCC® 49845™) was grown in MRS medium (Sigma-Aldrich) at 30 ºC in 5 % CO2 for 24 h. *Bacteroides thetaiotaomicron* (ATCC®29741™) was grown anaerobically at 37 ºC for 24 h in FAB medium (NEOGEN, Lancashire, UK). Bacterial cultures were harvested by centrifugation and suspended in PBS. A 1 ml aliquot of the suspension was used for to count colony forming units (CFU) by retrospective plating. The rest was resuspended in Neutral Buffered Formalin and left to fix for 18 h at RT.

Counting fixed bacterial cells. The cell suspension was counted using a bacterial counting kit for flow cytometry (Invitrogen). In brief, a 10% aliquot from the bacterial suspension was serially diluted to 1x 10^6 cells in 989 µl of NaCl. Bacterial cells were stained with 1 µl of SytoBC and 10 µl (1X 10^6) of counting beads were added to the suspension. Cells were counted in an LSR II Flow Cytometer (BD Biosciences). The acquisition trigger was set to side scatter and regulated for each bacterial strain to filter out electronic noise without missing bacterial cells. This value was approximately 800. The volume corresponding to approximately 2 x 10^7 CFU of each bacterial strain and 2.2 x 10^7 4T1 cells were mixed together.

Cell culture. *Mus musculus* mammary gland cancer cells (4T1) were grown at 37 ºC 5% CO_2, in RPMI-1640 (Sigma-Aldrich) media supplemented with 10% FBS (Sigma-Aldrich), 100 U/mL penicillin and 100 µg/mL of streptomycin (ThermoFisher), and counted with a NucleoCounter® NC-100™ (chemometect, Copenhagen).

Fixing cells in an agar matrix. An equal volume of sterile agar (1.5X of elution specified by the manufacturers) pre-aliquoted and kept at 56 ^{o}C, was pipetted into the cell suspension and thoroughly mixed by vortexing. The mixture was pipetted into a sterile cylindrical mould made from a 54 x 11 mm adapter tube (SARSTEDT, Cat No. 55.1570) and let solidify for 3 min. Once solidified, the disk was placed in 5 ml of formalin for an extra 24 h for 48 h fixation blocks or immediately processed for 24 h fixation blocks.

Dehydration and paraffin embedding of cell disk. Fixed cell disks were removed from the formalin and placed into a processing cassette. The cassettes containing the Protoblocks were dehydrated and paraffin embedded automatically with a LOGOS J (Milestone Medical, Bergamo). This protocol included 4 h dehydration with increasing concentrations of ethanol, clearing with 2 x washes of xylene and 3 x washes of isopropanol. Finally, the blocks were embedded in paraffin for 8 h and 32 min at 62 ^{o}C. . Once paraffinised, the Protoblocks' volume, diameter and height were measured with a calliper and by volume displacement [45]. Processed Protoblocks were placed in a 1.5 x 1.5 cm embedding mould and mounted to a processing cassette.

Sectioning. Blocks were sectioned keeping an aseptic technique either at 4 µm for imaging or at 15 µm for DNA purification. The cell load of each slide was calculated by dividing the total bacterial load by the volume of each slide.

Immunofluorescence and histochemistry. Cell integrity was evaluated with Gram staining (Sigma-Aldrich) or H&E staining with Mayer's haematoxylin (Sigma-Aldrich). Bacterial counts were confirmed in 3 sections stained with DAPI, 1:50 α-*E. coli* (Abcam, 137967), or 1:400 α-*S. aureus* (Abcam, 20920), and counterstained with either Alexa Fluor 488 (Jackson Immunoresearch Laboratories Inc., USA) donkey anti-rabbit Ig. Stained sections were mounted in ProLong Gold antifade reagent with DAPI (Invitrogen, UK). Gram-stained sections were counted in bright field using an Olympus BX51 microscope, with a 100X lens. Immunofluorescent stained slides were counted at 20X (4T1 cells) or 60X (bacteria) with a fluorescence microscope (Evos FL Auto). For each slide, at least 20 randomly selected fields of view were counted. The area of the field of view (FOV) was recorded using the microscope's software and used to calculate the volume counted.

2. DNA Analysis

DNA Purification. For purifying DNA from Protoblocks, unless specified, 10 x 15 µm sections aseptically collected sections were deparaffinated with 2X xylene washes and processed following procedures specified in the QIAGEN FFPE DNA kit protocol (Qiagen Inc., Valencia, CA, USA). DNA was eluted in Tris-HCL buffer and quantified with a Qubit™ dsDNA HS Assay Kit (Invitrogen, USA). For non-fixed bacteria, bacterial cultures were grown to an OD_{600} of 1. 2 ml aliquots were processed following procedures of the GenElute™ Bacterial Genomic DNA Kit Protocol with Lysozyme and Lysostaphin (Sigma) and eluted in 50 µl of Tris-HCl. In all cases, DNA was stored at -20^{o}C until further analysis.

Quantitative PCR. For quantitative qPCR, reactions were prepared using LUNA Universal qPCR (NEB, Ipswich, MA, USA) and 0.25 µM of each primer (sTable 1). The thermal profile included an initial denaturation of 1 min at 95 ^{o}C, and 40 cycles of denaturation at 95 ^{o}C for 10 sec, annealing for 15 sec at the primers' optimal temperature [54-56^{o}C] (specified by NEB's calculator for Hot Start Taq) and 20-40 sec of extension at 68 ^{o}C (20 sec for 200bp amplicons and 40 sec for 400-500 bp amplicons).

High-fidelity quantitative PCR reaction setup. Reactions were prepared using NEBNext-Ultra II Q5 Master Mix, 0.5 µM of each primer (sTable 1), 1.25 µM EvaGreen Dye (Biotium, CA, USA) and 37.5 nM ROX (Biotium, CA, USA) as a reference dye. The thermal profile included an initial denaturation of 30 sec at 98 ^{o}C, and 40 cycles of denaturation at 98 ^{o}C for 10 sec, annealing for 15 sec at the primers' optimal temperature [64-67^{o}C] (specified by NEB's calculator for Q5 High-Fidelity Master Mix) and 20-40 sec of extension at 72 ^{o}C (20 sec for 100 – 200 bp amplicons and 40 sec for 400 – 500 bp amplicons).

Quantitative qPCR assays parameters. Amplification was performed in an AriaMx (Agilent Technologies, USA) using DNA binding dye absolute quantitation experiment type. Each assay included triplicates of 5 points standards using log-dilutions of a 10^{7} copies gene block, designed upon a species-specific genetic region. Primers targeting these regions and maintaining a similar Tm (+/-2^{o}C) were designed

using the NCBI primer design tool and their parameters (ΔG, hairpins and dimers) verified using IDT's Oligo analyser tool. Primers and gene-blocks were acquired from IDT (Coralville, USA) (see sTable 1). qPCR efficiencies between 95% and 105% and R-square values higher than 0.995 were deemed as acceptable, all samples were ran in triplicate.

High-Resolution Melt (HRM) Curve Analysis. For melt curve analysis, it was essential to first normalise the amplifiable DNA fraction of samples tested. To achieve this, a quantitative qPCR was performed for fragments of the same length. The measured copy-numbers obtained by qPCR, were used to normalise the samples to 1 x 10^6 copies/µl. 20 µl reactions were prepared using 1X NEB Luna probe qPCR mix, 1.25 µM EvaGreen Dye (Biotium, CA, USA), 37.5 nM ROX as reference dye, 0.25 µM of each primer and 2.5 µl of copy-number normalised template DNA. *E. coli* primers rendering amplicons of 100, 200 and 500 bp were used for this assay (sTable 1). The amplification of the analysed target region was first amplified as specified for absolute quantitation, but included a final 2 min at 68ºC extension step. This was followed by high-resolution melt (HRM) analysis set to read fluorescence every 0.2 ºC with a 10 sec soak time from 65-95 ºC. All experiments were performed using an AriaMx thermocycler (Agilent Technologies).

Here, normalized fluorescence (Rn) obtained every 0.2ºC, across the temperature gradient (65-95 ºC) was used to monitor the melting temperature (Tm) profile of the template. Changes in the Tm profile are indicative of changes in the template sequence. To better observe this changes, the Tm profiles were plotted on a Tm difference (ΔTm) plot, where the Tm difference is represented by the deviation of the recorded Rn values of a Test plotted against those recorded for a non-fixed reference, for which the ΔTm is 0. Therefore, ΔTm = Rn Test – Rn of reference. Here where aberrant profiles that differ from the NF DNA with ΔTm <0.1ºC are typical of FFPE DNA, and are indicative of low-level, non-identical changes randomly distributed across the template [46]. Therefore, in these plots a lower ΔTm, is indicative of a reduced/lower number of sequence artefacts in the template. Raw Tm values were extracted from the AriaMx software and analysed in R environment, v3.4.4.

Sanger sequencing. Sanger sequencing was performed on 500 ng of purified and/or treated DNA for each replicate on the same genomic regions analysed by qPCR. Sequencing was performed by Eurofins Genomics.

WGS sequencing library preparation. For NF controls, DNA from bacterial cultures of *Escherichia coli* MG1655 and *S. aureus* Newman were grown as per section 1 to and OD_{600} of 1 and their genomic DNA purified using the GenElute™ Bacterial Genomic DNA Kit Protocol with Lysozyme and Lysostaphin (Sigma). For FFPE bacteria, DNA from Protoblocks containing either strain was purified using the QIAGEN FFPE kit plus specified treatment. In all cases DNA was eluted in 50 µl of Tris-HCl. Total purified DNA and/or repaired DNA was sent to GENEWIZ (Leipzig, Germany) where WGS was performed using 2 x 150 bp chemistry on an Illlumina HiSeq.

3. Optimising cross-link reversal

As described in Chapter 1, section 3, the product of the interaction of HCOH and biomolecules is the formation of crosslinks. These are ubiquitous in FFPE samples, and occur more frequently between dG and amino acids Lys and Cys in the form of DPCs [47, 48]. DPCs inhibit DNA amplification by blocking the processivity of DNA polymerases, terminating primer extension [49]. Despite their high prevalence in FFPE samples, it has been demonstrated that HCOH crosslinks are reversible. This reverse reaction is heat dependent [13], and can be assisted by pH, salt concentration and the incorporation of quenchers [50-52]. Heat treatment for decrosslinking, is essential for FFPE DNA purification and all available protocols and kits for FFPE DNA purification incorporate it, typically as a 1h incubation at 90^{o}C [14]. However, recent studies have found this high temperature incubation detrimental to DNA and shown that upon a reduction of temperature or time of decrosslinking, the appearance of sequence artefacts was reduced, although also reducing the amount of sequencing reads (decrosslinked DNA) [21, 24]. Thus, optimising the reaction conditions to allow lower incubation temperatures with equal decrosslinking yields, would reduce the adverse effects produced by high temperature incubation.

Temperature-point experiments. 10 x 15 µm sections from blocks loaded with 10^8 *E. coli* and *S. aureus* cells fixed for 24 h and stored for 3 months were distributed into 12 x 1.5 ml tubes. The deparaffinated and digested contents were pooled and distributed into 24 experimental replicates, 6 replicates per temperature point tested (90ºC, 80ºC, 72ºC and 65ºC). For temperature points 90 ºC and 80 ºC, incubation time was set for 1 h and for 72 ºC and 65 ºC it was set for 2 h. After decrosslinking, the DNA content was purified with the QIAGEN FFPE protocol.

Cross-link reversal buffer. Lysis buffers tested for crosslink reversal were TB1 (50 mM Tris-HCL (pH 8.0), 30 mM EDTA, 800 mM GuHCL, 0.5% Triton-X, 0.5% Tween-20), TB2 (50 mM Tris-HCl (pH 8.2), 100 mM NaCl, 1 mM EDTA, 0.5 % Tween-20, 0.5% NP40, 20 mM DTT) and TB3 (50 mM Tris-HCl (pH 8.0), 100 mM EDTA (pH 8.0), 100 mM NaCl, 1% SDS). 10 x 15 µm slides from blocks loaded with 10^8 *E. coli* and *S. aureus* cells fixed for 24 h and stored for 3 months were used per experimental replicate (6 per buffer tested). The samples were lysed and digested in the experimental buffer at 56 ºC for 1 h and decrosslinked at 80 ºC for 1h. After testing for decrosslinking buffers, an equal volume of buffer AL (column binding buffer) was added to the reaction and the DNA content purified following the QIAGEN FFPE kit protocol.

Verifying cross-link reversal strategy. A total of 10 x 15 µm slides from blocks loaded with 10^8 *E. coli* cells fixed for 48h and stored for 1 year were used per experimental replicate (6 per test). After decrosslinking, the DNA content purified with the QIAGEN FFPE kit.

4. DNA repair

Treatment with individual glycosylases. DNA purified from FFPE blocks loaded with 10^8 *E. coli* cells fixed for 24h or 48h was pooled and its concentration measured and normalised across tests. Aliquots with equal DNA concentration were used for each experimental replicate. All enzymes tested were acquired from NEB (Ipswich, MA, USA) and the verified enzyme activity provided by the supplier used to calculate amount of enzyme input. To calculate the enzyme input per ng of *E. coli* K12 MG1655 DNA, *E. coli* genomic data in sTable 2 was used.

For this, enzymatic activity was first normalised in terms of number of damaged nucleotides or lesions repaired by an enzyme unit in a standard 30 minutes reaction. An estimate of 0.05 – 0.1 % of damaged bases in FFPE DNA was used as a baseline. With this information, the number of damaged bases was first calculated per ng of DNA in the reaction and the enzyme units required to repair this damage. The units of enzyme used were optimised to fit the activity in a universal buffer and after titration experiments. The final units used in the reaction and the number of bases corrected per ng of *E. coli* DNA are listed in sTable 3. 40 µl reactions were set-up using a total of 400 – 1,000 ng of bacterial DNA. The reactions were run at 37 ºC for 30 min, after which enzymes were heat-inactivated with incubations specified in sTable 3. Treated DNA was cleaned using the Monarch PCR & DNA Clean-up Kit (NEB, USA). DNA concentration was measured with QUBIT (Invitrogen) and normalised DNA quantities analysed by quantitative PCR or HRM.

Assembling Base Excision Repair reaction.

Buffer: The BER pathway was reconstituted in a final buffer with 1X NEB CutSmart buffer (50 mM Potassium acetate, 20 mM Tris-Acetate, 10 mM Magnesium acetate and 100 µg/ml of bovine serum albumin, pH 7.9), supplemented with 100 µM of dNTPs, 50 µM of NAD+ and 2 mM of DTT. Enzyme efficiency in this buffer was analysed by comparing its activity with the buffer provided by the manufacturer. The compared enzyme activity was used to adjust the enzyme units used for the BER reaction.

Repair of excised bases: The repair of excised bases was accomplished with long (UDG) and short patch BER (FPG, Endo VIII), by incorporating the downstream enzymes that repair blocked ends (PNK) or AP sites (Endo IV), plus DNA polymerase and DNA ligase (sTable 4). The reactions were prepared with the buffer described above, using normalised DNA quantities and carried out at 37ºC for 30 min. The reactions were stopped with the addition of 2X volumes of Agencourt AMPure XP magnetic beads (Beckman Coulter, IN, USA) for DNA clean-up. Following the manufacturer's instructions, DNA was washed twice with 80% ethanol and eluted in 36 µl of Tris-HCl. DNA concentration was again measured for each reaction and

normalised DNA quantities were used for quantitative PCR, HRM, or by Sanger sequencing.

BER with combined glycosylases. These reactions were setup and carried out as described for BER reactions using only one glycosylase, with the difference that these reaction also included the downstream lesion repair enzymes (Endo IV and PNK) specified for the sub-pathway triggered by the glycosylases included. The reactions were analysed by HRM, Sanger sequencing and WGS.

5. Bioinformatics and statistical analysis

qPCR data analysis. Statistical analysis performed in the base R environment (v3.6.1). Visualisations were carried out using the ggplot2 package (v3.2.1).

Sanger sequence analysis. The effect of DNA repair enzymes on DNA sequence length and readability was assessed by Sanger Sequencing. The ratio of clipped sequence length to unclipped sequence length between samples was compared to elucidate this. Statistical analysis performed in the base R environment (v3.6.1). Visualisations were carried out using the ggplot2 package (v3.2.1).

WGS sequence analysis. All metrics relating to sequence data were calculated in the Linux environment, and using the QUAST tool (v5.0.2) and statistical analysis performed in the base R environment (v3.6.1). Visualisations were carried out using the ggplot2 package.

Method for variant calling:

Filtering: HiSeq sequence data was quality filtered. Only very high quality bases were considered to minimise the risk of sequencing errors causing false positive variants. Short fragments were also removed to reduce the likelihood of spurious alignments of regions from contaminant bacterial genomes. Trimmomatic (v0.38) was used to remove all reads shorter than 60bp in length, and to trim reads when the average per base quality in a sliding window of size 4 dropped below 30.

Alignment: Of the three possible Burrows-Wheeler alignment tools, the BWA-mem aligner was used as the average read length was 150bp, and BWA-mem (v0.7.17) is recommended when reads are over 70bp in length. Default settings were used with the exception of allowing alignments with a minimum score of 0, rather than the default 30. Given the stringent parameters used for read length and quality filtering, relaxing the minimum alignment score gave the best possible chance of variant detection. All samples were aligned to the original reference genomes.

Variant Calling: Variant calling was done with BCF tools, using the BCF call function. The variants were then filtered using the norm and filter functions within BCF tools. Filtering was done to remove variants when the read depth was below 10, the quality was below 40, or when the variant identified was not supported by both the forward and reverse read of a read pair. The number of variants identified was then normalised between samples based on the read coverage in the initial alignment BAM file.

Validation: Using the Picard tool within the GATK suite, all samples were down-sampled to ensure SNP: Coverage ratio remained constant when coverage was reduced to lowest level present in samples.

RESULTS

1. Characterisation of bacterial FFPE DNA damage

Measuring fragmentation of PCR readable DNA: The length of PCR-readable fragments from bacterial DNA subjected to FFPE treatment was measured by quantitative PCR. Targeting a 525 bp chromosomal region, primers were designed to amplify DNA fragments of lengths 200bp, 300 bp, 400 bp and 500 bp. Template DNA was purified from FFPE blocks loaded with 1 x 10^8 E. coli cells, fixed for 48 h and stored for > 6 months. Each qPCR reaction was loaded with 5 ng of DNA, corresponding to 1 x 10^6 CFU. As seen in Figure 1a, the quantity of amplifiable DNA is significantly reduced after FFPE treatment. For non-fixed (NF) DNA, the amplification of PCR-readable fragments is almost 100 %, and is independent of fragment size, whereas a log-fold reduction of amplifiable DNA is observed for even short (200 bp) fragments of FFPE DNA ($p < 0.001$). Importantly, this becomes more pronounced as fragment length increases, with significant correlation between reduction in the quantity of amplifiable DNA and fragment length, leading to a log-fold reduction in amplifiable DNA quantity between 200 bp and 500 bp fragments ($p < 0.001$).

Assessing the extent of formaldehyde cross-links in FFPE bacterial DNA: The presence and frequency of formaldehyde crosslinks present in bacterial DNA was assessed by comparing the quantity of amplifiable DNA obtained after performing or omitting a crosslink reversal incubation on paired-samples (n = 6),), a strategy resembling the straightforward FAIRE method [13]. As can be seen in Figure 1b, crosslinking was evident regardless of fragment size, with an 18.5 (500 bp) – 30 (200 bp) fold increase in amplifiable DNA observed after crosslink reversal, indicating that 95% –97% of the amplifiable DNA in the sample held crosslinks that inhibited its amplification.

Evaluating the presence of damaged nucleotides: The presence of damaged bases in bacterial FFPE DNA was investigated by subjecting FFPE-DNA to the activity of DNA glycosylases targeting base oxidation, deamination and carboxylation with enzymes listed in sTable 3. DNA lesions resulting from DNA glycosylase activity (AP

sites and 3'P) [36, 39], inhibit amplification [53]. Therefore, DNA glycosylase activity can be measured by comparing the quantity of amplifiable DNA in a sample after treatment/no treatment with a DNA glycosylase, with a decrease in amplification implying the presence of the targeted DNA damage. As seen in Figure 1C, a decrease in amplifiable DNA was noticeable in concentration normalised samples after treatment with all glycosylases, with the highest activity observed for UDG and FPG as indicated by the 35% – 50% and 67 – 80% reduction in the recovery of PCR readable DNA fragments after treatment ($p < 0.001$) (Figure 1c). It should be noted that Endo VIII activity is not measurable by this PCR analysis, as lesions targeted by this enzyme (hydantoins) are PCR inhibitory, thus, the removal of this damage would not have any effect on the amount of amplifiable DNA template. [54].

Assessment of DNA sequence quality by sequencing: Overall DNA damage is reflected in the outputs of sequencing. Damaged bases and single strand breaks present as sequencing misreads, such as chimeras, indels and SNPs that lead to poor quality reads, which will be routinely filtered out prior to analysis. As seen in Figure 1d, a significant decrease in high-quality, sequencing-readable DNA was observed in both Sanger sequencing and WGS, for FFPE samples compared with their paired NF samples. This was accentuated by prolonged DNA fixation, where the reduction of high quality sequences reaches 30% ($p <0.001$).

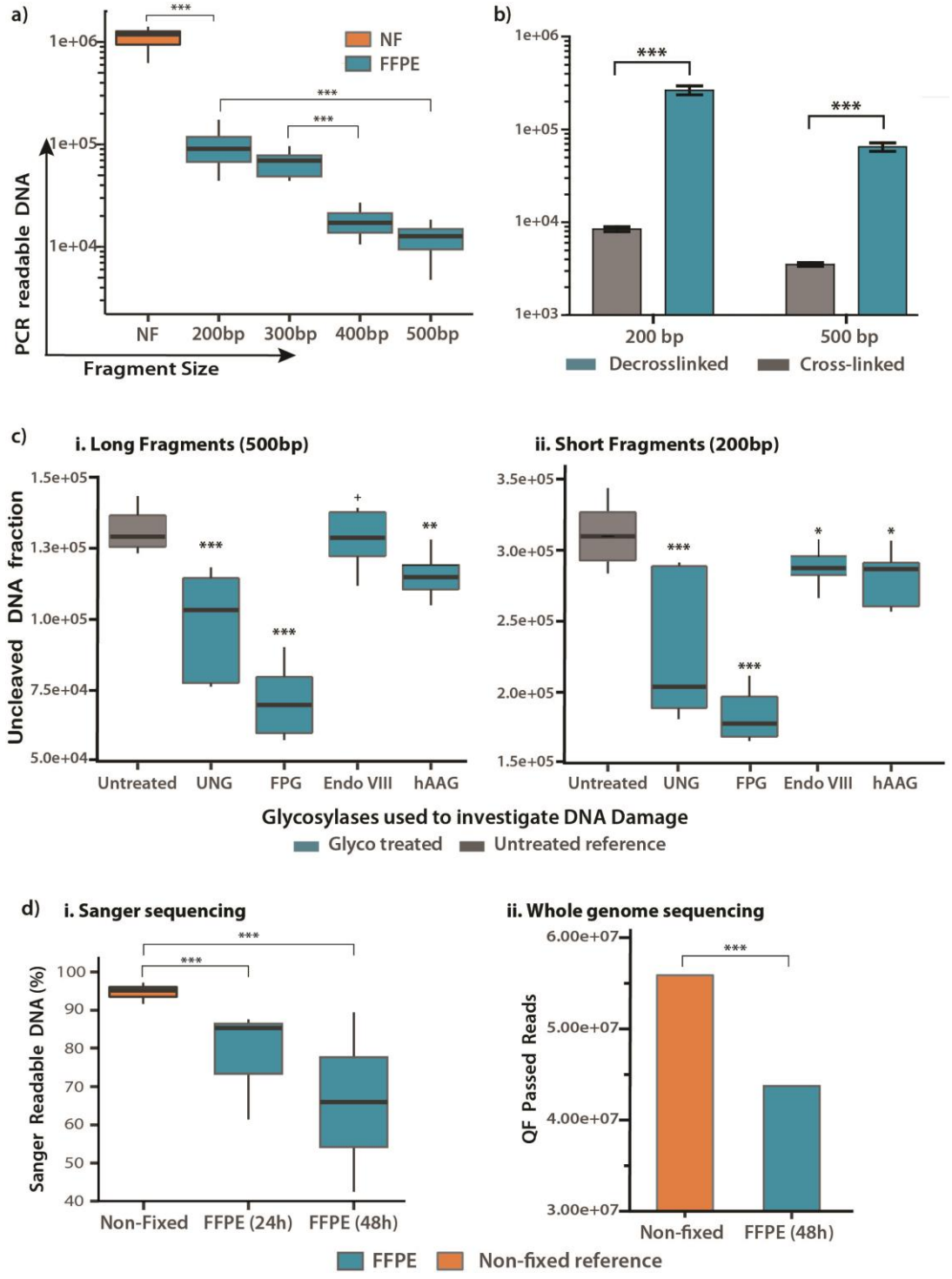


Figure 1. Analysis of DNA damage.

a) Measuring fragmentation of PCR-amplifiable DNA. *For NF bacteria, amplification of all fragment lengths was equal and grouped in the same box (n = 28). For FFPE bacteria (n = 24 for each box), a linear fragment-length correlation is evident, with a log-decrease observed from NF to FFPE 200bp fragments and a log-decrease between short (200bp) and long (500 bp) fragments (P < 0.001).*

b) Assessing the extent of cross-links in bacterial DNA. *DNA from FFPE blocks containing E. coli cells was subjected (n = 6) or not (n = 6) to a high temperature crosslink reversal treatment. The bar-plot shows the quantity of amplifiable DNA obtained +/- crosslink reversal for long and short DNA fragments. Without decrosslinking, only 3 - 5% of the available DNA template is amplifiable for PCR.*

c) Evaluating the presence of damaged nucleotides via glycosylase treatment. *Box plots show the quantity of amplifiable DNA post treatment with the respective glycosylase (n = 6 in all cases).*

d) Assessment of DNA sequence quality by sequencing. *(i) Sanger sequencing showing the percentage DNA falling within the high-confidence read region for each sample. (ii) Whole genome sequencing showing the number of quality filter pass reads for FFPE and NF bacteria.*

2. Development of a DNA repair strategy

Having characterised the nature of FFPE-induced damage to bacterial DNA, an appropriate repair strategy was devised, as outlined in Figure 2.

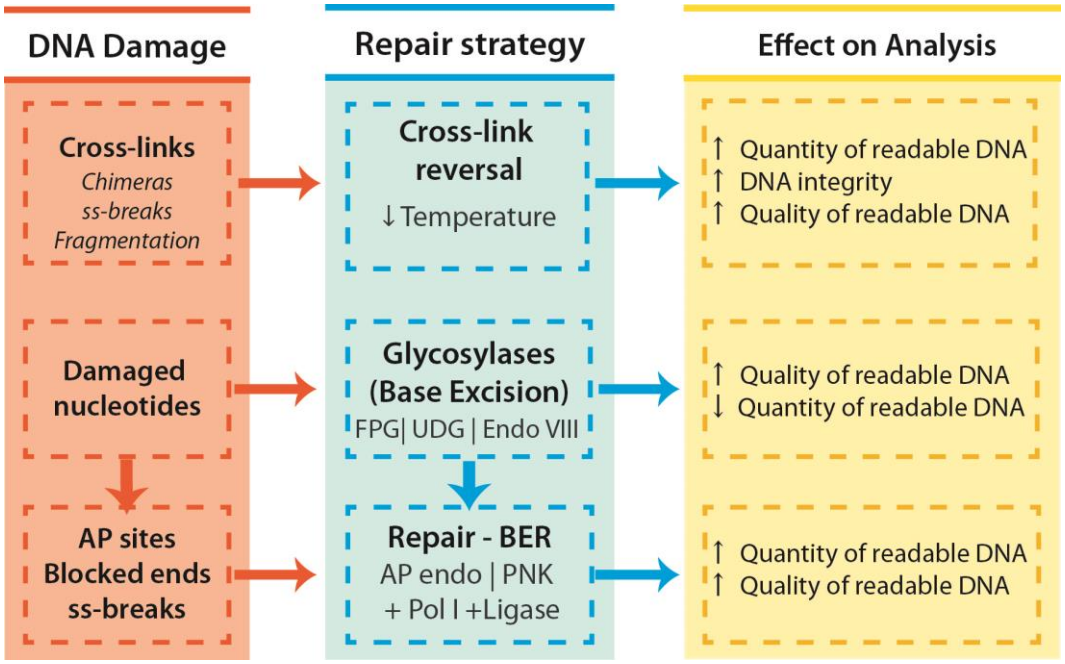


Figure 2. Summary of strategies applied for improving integrity, quantity and quality of bacterial DNA derived from FFPE samples.

(a) Exposure of DNA to denaturing temperatures (90 °C) aids decrosslinking, but increases the rate of depurination and ss-break events that lead to the formation of ss-DNA regions known to favour the misincorporation of nucleotides (A – rule) or generate sequence chimeras. Therefore, milder decrosslinking reactions will reduce the rates of these occurrences.

(b) The FFPE process damages DNA bases. The removal of damaged bases by glycosylases improves the quality of readable DNA by removing from the PCR pool damaged template that would otherwise lead to misincorporation of bases leading to SNPs. The product of either glycosylase treatments are AP sites (UDG) or 3' blocked ends (FPG, Endo VIII) that block polymerase activity.

(c) These blocking artefacts are repaired by either an AP endonuclease (AP sites → Endo IV), leaving a 3'OH and 5'dRP, or a Phosphokinase (3'P → T4 PNK), leaving a 3'OH and a 5'P. Only when ends are repaired (3' OH and 5' P / 5'dRP) is the DNA repair polymerase (Pol I) able to incorporate nucleotides that are subsequently sealed with a high fidelity DNA ligase (*E. coli* DNA ligase).

Optimisation of decrosslinking

Crosslinks block polymerase processivity, reducing yields of PCR readable DNA [49]. Formalin induced crosslinks are reversible upon heat exposure and all available FFPE DNA preps include a high-temperature (decrosslinking) incubation step [13]. Recently, it has been shown that this incubation, despite improving PCR yields, reduces DNA sequence quality and fragment length [21, 24], making it unsuitable for microbiome research of FFPE samples. For this reason, we aimed at investigating strategies that reduce heat-exposure in order to find the optimal balance that improves the output DNA sequence quality without significantly affecting its yield.

Temperature: The effect of decrosslinking temperature on the yield of amplifiable DNA was investigated by quantitative PCR in DNA extracted from FFPE blocks loaded with *Staphylococcus aureus* (Figure 3ai) and *E. coli* (Figure 3aii), fixed for 24 h and stored for 3 months. Reactions were loaded with 10^6 copies of template and incubated at 90 ^oC for 1 h (reference protocol = industry standard mammalian DNA isolation from FFPE tissue), 80^oC x 1 h, 72 ^oC x 2h or 65 ^oC x 3 h. Compared with the reference 90 ^oC (QIAGEN protocol), no significant difference in amplification of PCR readable DNA was observed at 80 ^oC for both bacteria ($p > 0.05$), while a 4X (*E. coli*) and a 10X (*S. aureus*) decrease in the amount of PCR readable DNA was evident at both 72 ^oC and 65 ^oC ($p < 0.001$). In this case, PCR amplification is indicative of the template fraction that was efficiently decrosslinked.

Buffers: The ability of three protein lysis buffers (also used for protein digestion) in setting conditions (pH, ionic strength, enthalpy disruption) that facilitate

decrosslinking at 80oC were examined: Test Buffer 1 (TB1) – based upon the protein denaturing properties of chaotrophic agents (Guanidine hydrochloride); Test Buffer 2 (TB2) – Denaturing proteins with a reducing agents (DTT); Test Buffer 3 (TB3) – relying on the denaturing properties of an ionic detergent (Sodium dodecyl sulphate). Decrosslinking with the three buffers was tested against the reference buffer (Buffer ATL, Qiagen FFPE Kit) at 80 oC x 1 h. The effect of each buffer upon decrosslinking efficiency was assessed quantitatively by comparing the quantity of amplifiable DNA recovered after treatment. Contents of FFPE slides loaded with *E. coli* and *S. aureus* cells were suspended in each buffer (n = 6). Purified DNA was subjected to qPCR for amplification of a 500 bp fragment. TB1 and ATL buffer displayed the highest yield ($p > 0.05$), significantly higher than TB2 ($p < 0.05$) and TB3 ($p < 0.01$); (Figure 3b).

Evaluating DNA sequence quality of optimised strategy: The optimised strategy 1 h at 80 ^{o}C in TB1 was tested against the standard protocol 1 h at 90 ^{o}C in QIAGEN ATL Buffer for its capacity to decrosslink DNA, indicated by the yield of 500 bp PCR products (Figure 3ci), and the sequence quality of the fragments yielded (Figure 3cii, iii). This was tested in DNA sourced from FFPE blocks loaded with *E. coli* fixed for 48 h and stored for 1 year (representing maximum damage conditions). For quantitative analysis, reactions were loaded with normalised DNA concentration. For qualitative analysis, reactions were loaded with 10^6 amplifiable copies of the DNA fragments. As shown in Figure 3c (i), with the new strategy, the yield of amplifiable DNA did not differ significantly from that of the QIAGEN protocol. However, the sequence quality of DNA recovered was improved with the new strategy. As seen in Figure 3c (ii), the melting temperature (Tm) of samples treated with the new strategy was less variable and closer to that of paired-NF DNA, exhibiting a Tm difference [ΔTm (%)] of 2.82 (not significant), versus 3.02 ($p < 0.05$) for the QIAGEN protocol. This was further explored with HRM (detailed in methods), where aberrant profiles (from that of NF DNA) are indicative of sequence aberrations typically found FFPE DNA [46]. ΔTm plots in Figure 3ciii, show that the ΔTm for samples decrosslinked with the new strategy (ΔTm (%) = 3.5) is significantly lower than that of the QIAGEN protocol (ΔTm (%) = 6.1) ($p < 0.05$). This indicates that with the new strategy, without compromising DNA yields, the sequence quality of decrosslinked template is less damaged (resembles more NF DNA).

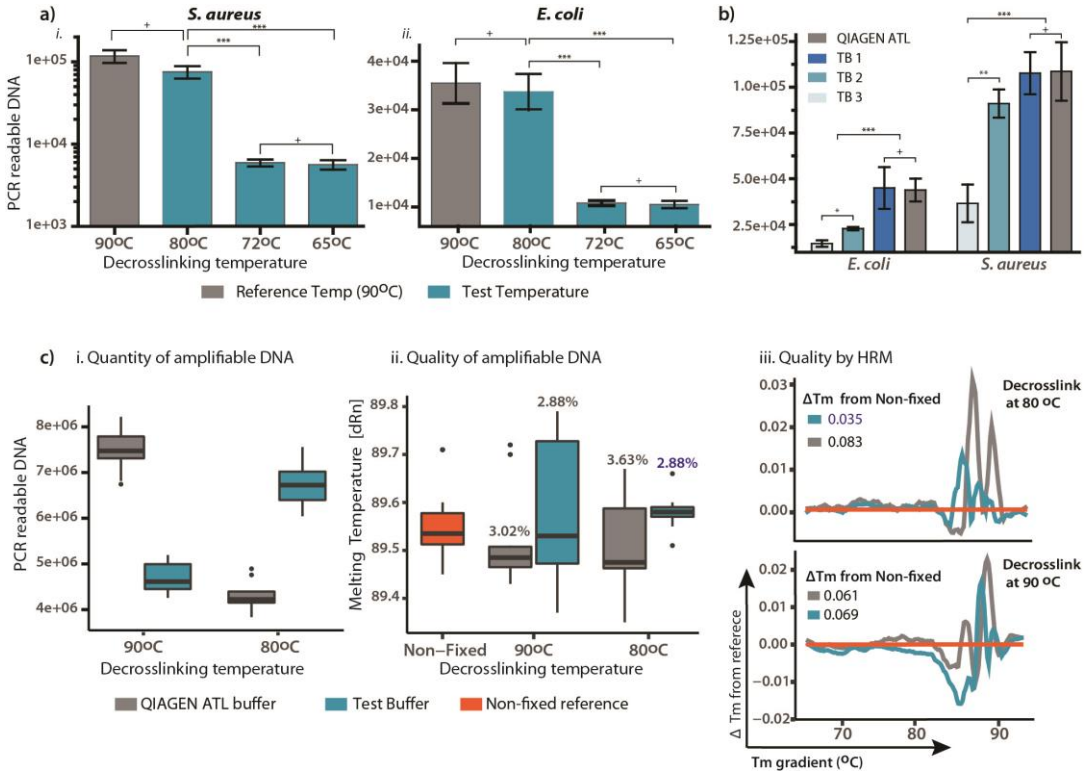


Figure 3. Optimising a decrosslinking strategy.

a) Temperature. *The bar plots shows the recovery of 500 bp PCR readable DNA fragments after testing 3 crosslink reversal incubations (blue, for each bar n = 6) against a reference (90ºC) incubation (grey, n = 6).*

b) Buffer. *Three buffers were tested against the reference buffer (ATL) at a 90ºC x 1h incubation. The amount of amplifiable DNA measured by qPCR of a 200 bp fragment in E. coli and S. aureus (for each bar n = 6).*

c) Evaluating the optimised strategy. *The quantity of amplifiable DNA (i) and the sequence quality of DNA (ii, iii) was assessed for a 500 bp DNA fragment. The performance of the optimised protocol (blue) was measured by comparing with the reference protocol (90ºC with ATL buffer) (grey). Box plot (i) shows the absolute quantity of amplifiable DNA from template DNA with normalised concentration (n=6 for each box). In box plot (ii), the Tm of the tested conditions (n=6 for each box) is compared with that of NF DNA (orange, n =6). The Tm difference (ΔTm) between the test and NF DNA is indicated above each box. (iii) HRM plot – ΔTm of tests plotted against the Tm of NF sample (orange), with average ΔTm from NF shown above each plot (n= 6 for each line).*

DNA glycosylases reduce sequence alterations in FFPE DNA

After examining their activity on FFPE DNA (Figure 1c), the effect of treatment with DNA glycosylases on DNA sequence quality was assessed by: a) Tm analysis, b)

Sanger sequencing, and c) HRM. For Tm analysis and HRM, all reactions were loaded with 1 x 10^6 genome copies of DNA sourced from FFPE blocks loaded with E. coli and set to amplify 3 x 100 bp fragments (Figure 4a and sFigure 1). For all the regions analysed, the Tm of samples treated with glycosylases significantly changed from FFPE untreated samples ($p < 0.001$) and came closer to resemble that of the NF reference. This was further assessed by HRM, by comparing the melting profile of a 200 bp fragment (as explained in figure 3 and methods). As seen in Figure 4c, the plotted ΔTm (from paired-NF) of glycosylases treated FFPE DNA was found to be much lower than that of untreated FFPE DNA. The same effect was evident with Sanger Sequencing (Figure 4b), where treatment with DNA glycosylases significantly improved ($p < 0.001$) the number of high-quality reads recovered, increasing the readability of DNA to levels no longer significantly different from NF DNA.

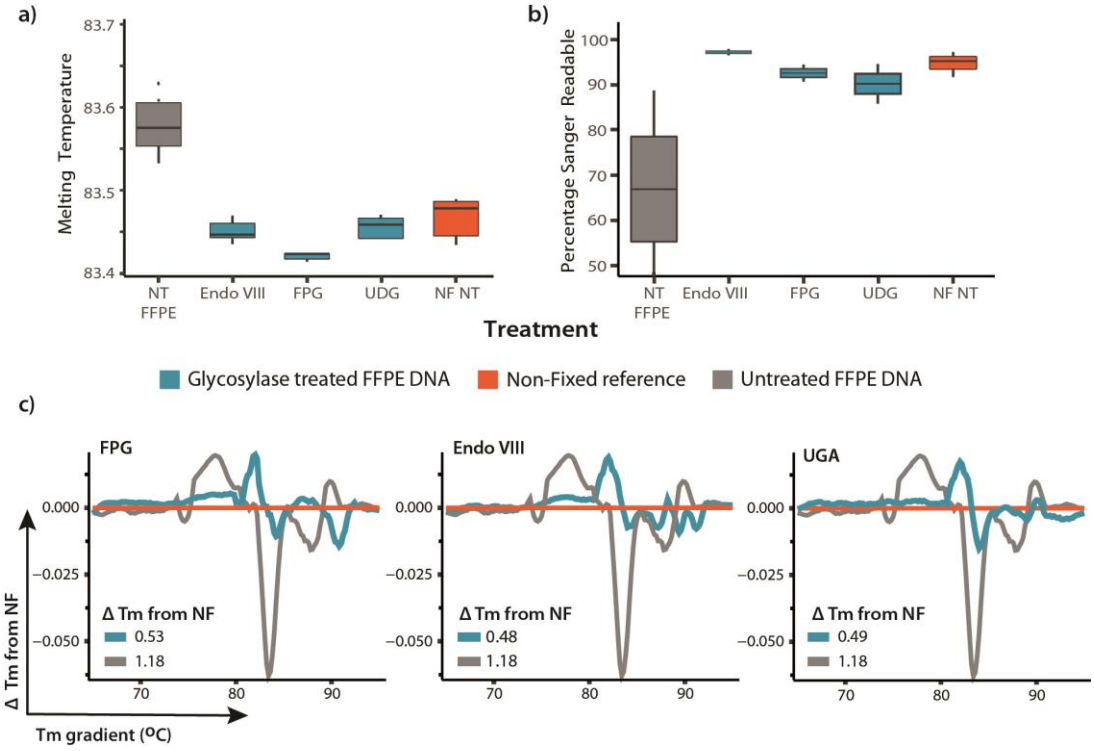


Figure 4. DNA glycosylases reduce sequence alterations in FFPE DNA. The reduction of sequence alterations in FFPE DNA (fixed for 48h) by treatment with the selected glycosylases was confirmed by: ***a) Analysis of their melting temperature (Tm)*** (n = 6 for each box). ***b) Sanger sequencing readability*** (n = 3 for each box). ***c) HRM*** (n = 6 for each line). In all tests performed, treatment with DNA glycosylases improved the amplifiable sequence quality. Grey: untreated FFPE samples. Orange: NF reference. Blue: glycosylases.

Development of an in vitro Base Excision Repair system

For the *in vitro* reconstitution of the BER pathway, a suitable universal buffer was sought and tested by examining enzymatic activity for each enzyme (see Methods) and compared with activity in their recommended buffer (see sFigure 2). Optimisation of enzyme and co-factor quantity usage was then performed (sTables 3 and 4).

First, the BER pathway was reconstituted for single repair pathways triggered by a single DNA glycosylase, with units and enzymes listed in Table 2 and 3, and its performance tested by HRM analysis. Figure 5a shows the HRM plots of DNA exposed to the BER pathway reconstituted for FPG, UDG or Endo VIII. As explained in methods, the more similar a DNA sequence is to the NF reference, the lower the difference in melting temperature (ΔTm closer to 0). As seen in Figure 5a, exposure of DNA to each reconstituted BER pathway led to a reduction in ΔTm in FFPE DNA and an increase in the quantity of PCR readable template (sFigure 3) suggesting a reduction in the frequency of sequence artefacts. The frequency of sequence artefacts observed after treatment was more effective for the FPG driven BER reaction, with a ~50% decrease in ΔTm observed for untreated samples, this was followed by Endo VIII with a ~31% reduction and finally UDG with a ~14% decrease in the ΔTm. These results indicate that BER was reconstituted correctly and that these reconstituted pathways effectively corrected sequence artefacts without reducing the PCR readable template.

Subsequently, the reconstitution of a BER system able to target different types of DNA damage found on FFPE samples was addressed by mixing the pathways for the glycosylases treated in the system. Since FPG-BER (Figure 5a) yielded the best results for single glycosylase-BER reactions, this enzyme was combined with ENDO VIII and/or UDG and their efficiency in reducing sequence artefacts tested by HRM. As shown in Figure 5b, all combinations resulted in sequences with ΔTm lower than those of untreated FFPE DNA. The FPG + UDG mix showed the best performance at reducing the ΔTm (31 %), followed by FPG + Endo VIII (18 %). However, in terms of improving the PCR readability of a 500 bp fragment, FPG + Endo VIII (47% increase, $p < 0.01$) outperformed FPG + UDG (30% increase, $p < 0.01$), as measured by Taq qPCR. This effect was confirmed by high-fidelity qPCR (providing a more

stringent discrimination of damaged and repaired sequence), where FPG + UDG showed a 20% increase and FPG + UDG only a 4% increase of amplifiable DNA (sFigure 4). To confirm these results, a normalised DNA quantity from 6 replicates for each BER mix and 6 unrepaired samples were pooled into one (n = Σ6) and sent for analysis by WGS (Figure 5c). At this level of resolution, it is evident that the repair mix with FPG + Endo VIII offered the highest improvements in sequence quality in terms of providing *(i)* a coverage 4X higher than unrepaired, *(ii)* 4X more total reads and quality filter (QF)-passed reads, and *(iii)* a 50% reduction in the number of variants detected per sequence coverage. This repair mix was thus selected as the best repair mix for bacterial FFPE DNA.

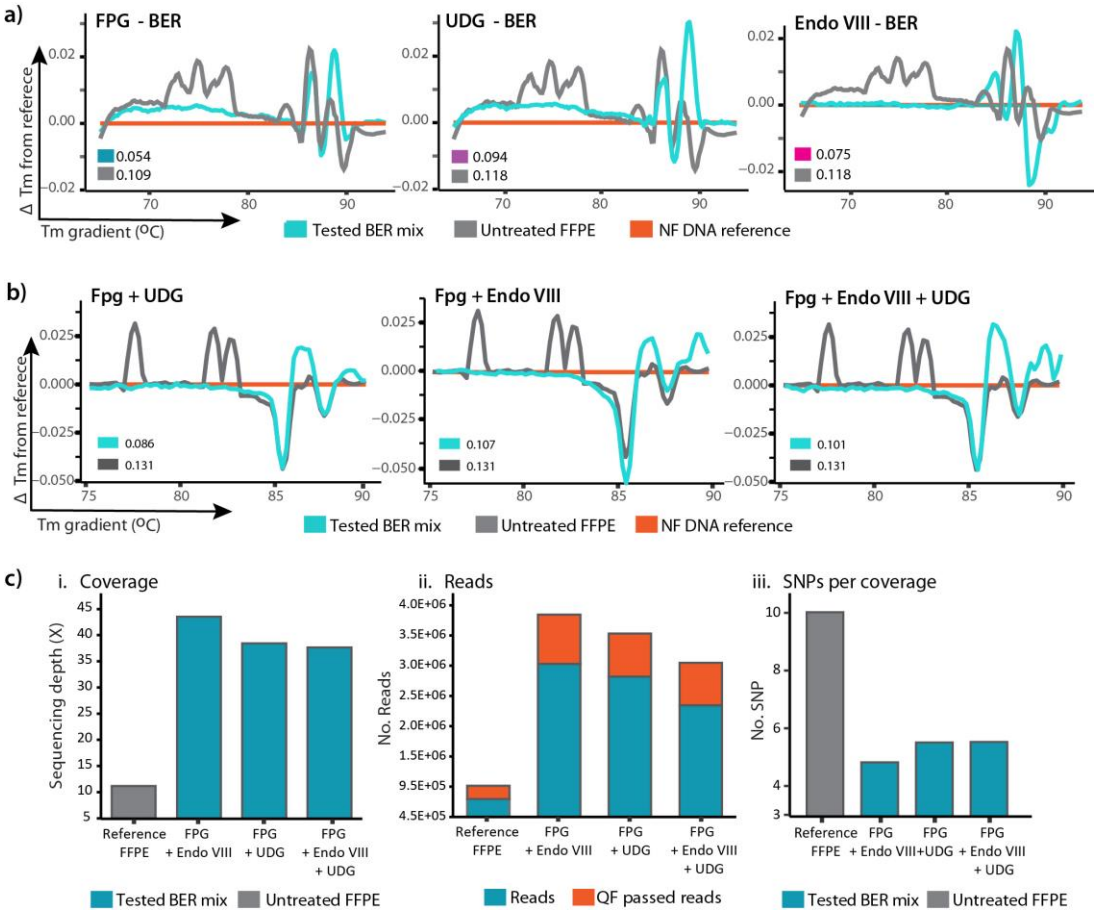


Figure 5. Reconstitution of BER pathway repairing FFPE DNA damage.

a) Single glycosylase BER. *The BER pathway was reconstituted first as single pathways triggered by either UDG, FPG or Endo VIII. The efficiency of each system in correcting DNA damage was tested by HRM (n = 7 for each line). The more similar a DNA sequence is to the NF reference, the lower the difference in melting temperature (ΔTm closer to 0). FPG showed the highest efficiency in correcting FFPE DNA damage as evidenced by the lowest ΔTm of*

0.054. ***b) Multiple glycosylase BER****. Mixes containing FPG show improved sequence quality as evidenced by reduced ΔTm vs untreated.* ***c) WGS****. To further confirm these results, 6 replicates treated with each mix were pooled (n = Σ6) and analysed by WGS. Data validated that all mixes improved the sequence (i) coverage, (ii) number of reads and QF passed reads and reduced the amount of SNPs (iii). The best performance in all cases was observed in the BER mix with FPG and Endo VIII.*

Analysis of combined decrosslinking and BER treatment

The sum of the above treatment strategies (decrosslinking and DNA repair), was tested by WGS in DNA sourced from FFPE blocks containing a mix of 5 bacterial strains, fixed for 48 h and stored for 2 months. DNA was decrosslinked at 80 ºC with TB1 (methods) and repaired with the FPG + Endo VIII-BER repair mix. The results of this were compared with those obtained from paired-samples treated with the reference protocol (decrosslinking at 90 ºC with QIAGEN ATL buffer, without DNA repair), and NF DNA obtained from equal cell contents. Experimental replicates were pooled (n = Σ6) and sent for WGS analysis. Results for this analysis are shown in Figure 6 and sFigure 5. The results obtained from exposing bacterial FFPE DNA to the proposed new protocol indicate that bacterial FFPE DNA treated with the proposed method shows an improvement in integrity, readability, and sequence quality, as evidenced by: (i) Integrity [Average fragment length (a, b)]: Plotted in Figure 6a, are the average fragment lengths measured with a fragment analyser. Fragment length of DNA treated with the new protocol (444 bp) is 3.3X longer than that treated with the reference protocol (decrosslinking at 90 ºC with QIAGEN ATL buffer, without DNA repair) (136 bp). Importantly, this raises the average fragment length to that of fragments typically desired for 16S sequencing (460 bp). The same effect was observed in the length of fragments read by WGS, where fragment lengths were 2-3 bp longer on average (Figure 6b). (ii) Readability: With the new protocol, the number of Total Reads and (QF)-pass reads per layer of coverage were increased by 24% and 34% respectively, and the ratio of QF-passed to Total reads increased by 8.4%. (iii) Sequence quality: This was measured in terms of number of sequence artefacts detected. The number of chimeric reads per coverage detected in samples treated with the new protocol was reduced by 57 % (p = 0.37) (Figure 6e). Similarly, the number of SNPs detected was reduced by 58% (p = 0.41) (Figure 6f and sFigure 5) in all strains tested. Despite the reduction in SNP's being uniform across all strains tested, FFPE

was found to produce a different SNP profile in Gram positive bacteria vs Gram negative bacteria (sFigure 6), which warrants further investigation.

Similar improvements in DNA quality and quantity to those shown in bacterial DNA were also obtained for the mammalian cell line used (4T1), where a 21% decrease in the amount of SNPs per layer of genome coverage and a 65% increase in the breadth of genome coverage was observed in the DNA treated with the proposed method (Figure 7). All of these findings are coherent with results from quantitative PCR and Tm analysis. Although these improvements are not supported by statistical significance, given the considerable effect size, we are confident that this lack of significance is due to sample size alone. Altogether, the sum of strategies proposed here were thoroughly investigated by PCR/sequencing. These results consistently indicate an improvement in the sequence integrity, readability and quality of readable bacterial FFPE DNA.

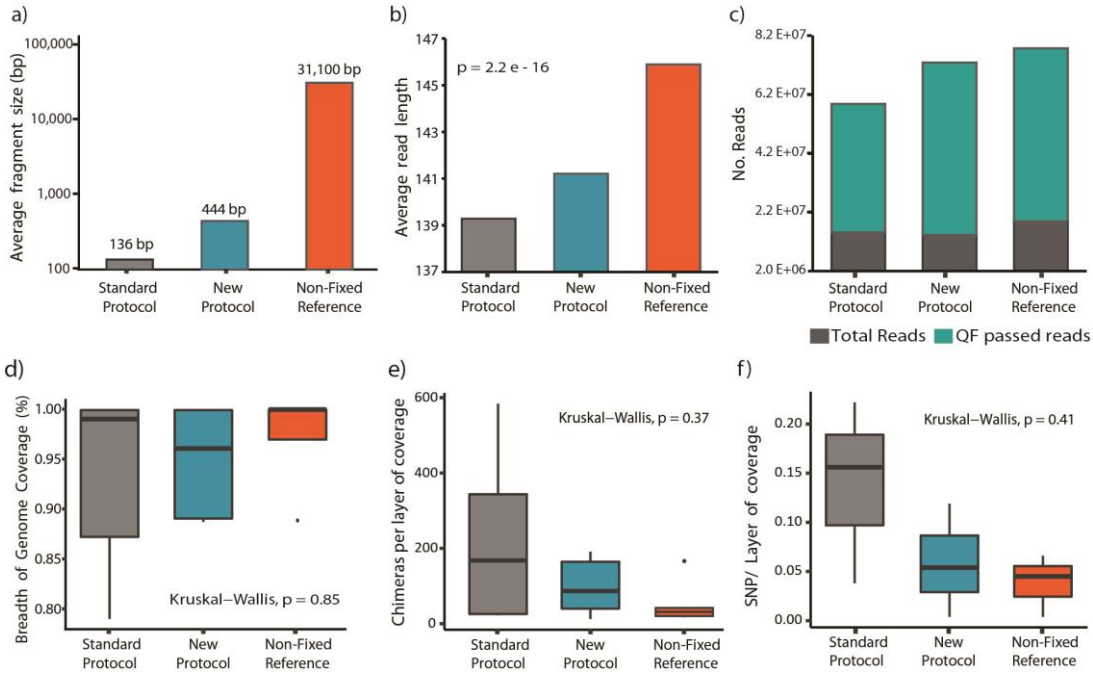


Figure 6. Combined protocol – bacterial DNA.

Outputs of Bioanalyser and whole genome sequencing for bacterial FFPE DNA exposed to the combined treatment (blue, labelled as New Protocol, Σn = 6). This was compared with that obtained from 6 pooled paired-samples decrosslinked with the reference protocol (90ºC, ATL) and unrepaired (grey, Labelled reference protocol, Σn = 6) and that from DNA obtained from NF samples with the same bacterial and DNA content (orange, Labelled NF, Σn = 3). Improvement in DNA readability, sequence quality and integrity was measured by: ***Integrity***

(fragment length): ***(a) fragment analyser (b) WGS. Readability****: (c)* ***Quantity of reads and filter pass reads*** *per coverage.* ***(d)*** *% Breath of genome* ***coverage****.* ***Sequence quality****:* ***(e) Number of chimeric reads*** *per layer of coverage.* ***(f) Number of SNPs*** *per layer of coverage.*

Improvements in DNA quality and quantity were also obtained for mammalian DNA (4T1), where a 21% decrease in the amount of SNPs per layer of genome coverage and a 65% increase in the breadth of genome coverage was observed in the DNA treated with the described method, although not supported by statistical significance.

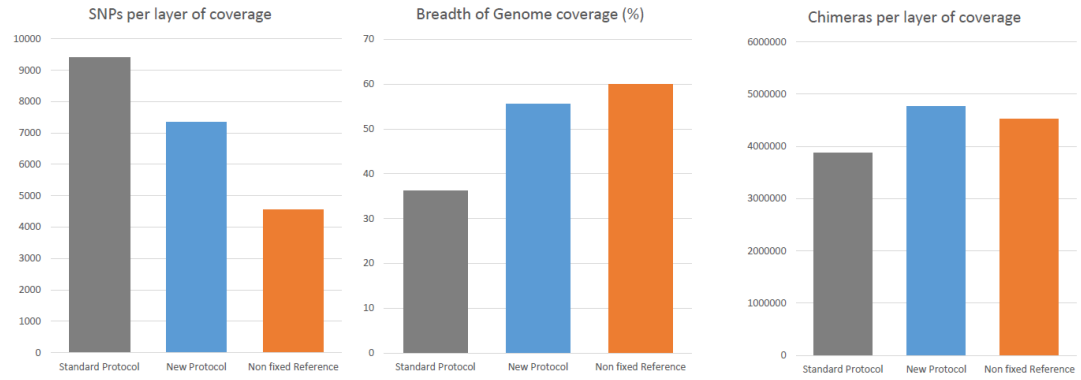


Figure 7. Sequence artefacts in mammalian DNA.

DISCUSSION

To our knowledge, this is the first such study in prokaryotic DNA, where an understanding of effects of FFPE on DNA, and impact on downstream analyses is arguably even more important. Our results show bacterial FFPE DNA to be a poor PCR template, with a log-fold reduction in the recovery of DNA fragments. This can be at least partially attributed to DNA fragmentation, since an inverse correlation between fragment size and PCR readability was shown (Figure 1a), culminating in a log fold reduction in recovery between 200 bp and 500 bp fragments.

Crosslinks were found to be ubiquitous in FFPE bacterial DNA (Figure 1b), and potentially more prevalent than in FFPE human DNA, based on previous research [12, 24]. Current decrosslinking protocols have been found to induce sequence alterations [21], and reducing heat-exposure has been proposed to prevent this damage [21, 24]. Our results are in agreement with these hypothesis, as a reduction from 90 ^{o}C (current protocols) to 80 ^{o}C, showed a significant reduction in off-target effects, without compromising the decrosslinking efficiency. Here, we hypothesise that TB1 (containing 50 mM Tris-HCL (pH 8.0), 30 mM EDTA, 800 mM GuHCL, 0.5% Triton-X, 0.5% Tween-20) established reaction conditions that promoted decrosslinking at a lower temperature. This could be explained by a higher protein denaturing capability of GuHCL (facilitated by a higher Proteinase K activity) [55-57], but also because GuHCL reduces the Tm of DNA (while maintaining high hybridisation stringency) [58, 59]. This would facilitate the exposure and hydrolysis of ubiquitous DNA-Protein crosslinks [51, 60] and DNA-DNA complexes [61-63] at lower temperature [64], reduce potential straining of the DNA structure, and maintain a high base paring fidelity. Although this could have also been assisted by other reaction conditions (such as pH and ionic strength) [61, 65-67], Tris-HCl formaldehyde scavenger activity [50, 51] or possibly Guanidium-formaldehyde interactions, this requires further investigation.

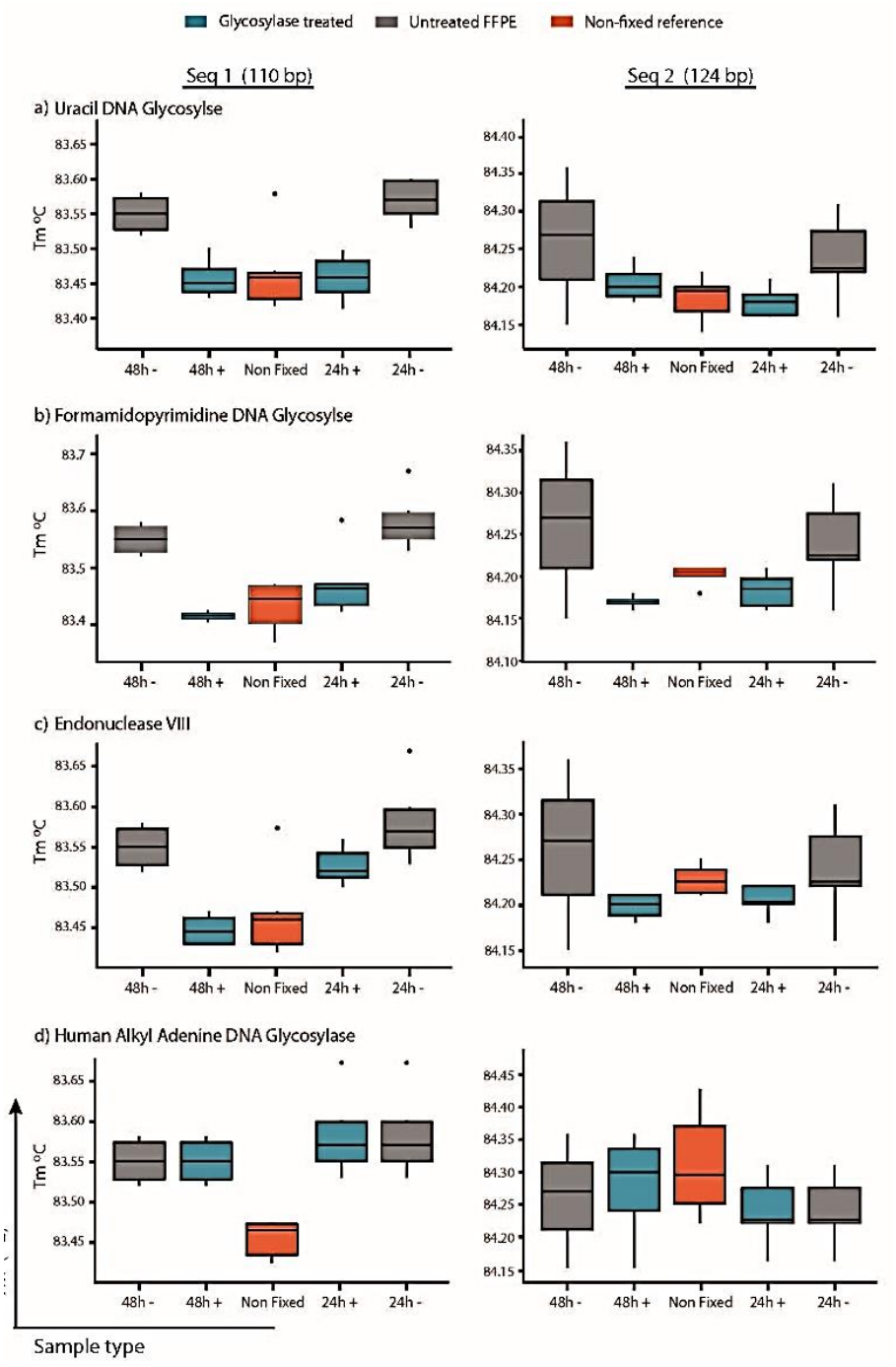
Treatment with glycosylases significantly reduces the appearance of sequence artefacts in FFPE DNA. Glycosylases generate blocked ends that are in most cases, unsuitable for amplification. This effect was confirmed in all glycosylases tested. Studies performed in human DNA have shown that cytosine deamination to uracil is

the main source of sequence artefacts in FFPE DNA [12]. However, this has been found controversial [14, 21, 24, 68]. Our data suggest that DNA damage found in bacterial FFPE DNA is primarily driven by oxidation and subsequent cytosine deamination, as evident in higher activity observed for FPG and Endo VIII. It is-known that oxidised products of cytosine can trigger deamination [69]. While UDG is able to repair some of the oxidised deaminated lesions (5-OH dU), Endo VIII has a broader spectrum of target products of oxidation and deamination. Quantitative and qualitative analysis by qPCR (Figure 5a, sFigure 4) and sequencing (Figure 5c) of samples treated with Endo VIII BER consistently yielded better results than UDG BER did, in terms of template readability and sequence fidelity. Interestingly, samples treated with Endo VIII alone showed an improved sequence quality. Given that damage targeted by Endo VIII is PCR inhibitory, this might be indicative of activity in non-blocking lesions (Fapy-A), reflect PCR errors triggered by blocking lesions (jumping PCR), or be due to a reduction of Taq Polymerase fidelity (A rule and/or deletions) [70, 71]. While the HRM melting curve analysis provided a valuable guide, confirmation was provided by qPCR and sequencing data. After exhaustive comparisons, the strategy found to be most effective involves decrosslinking using a chaotrophic agent at 80 ^{o}C, followed by DNA repair using a combination of Formamidopyrimidine DNA glycosylase and Endonuclease VIII.

CONCLUSION

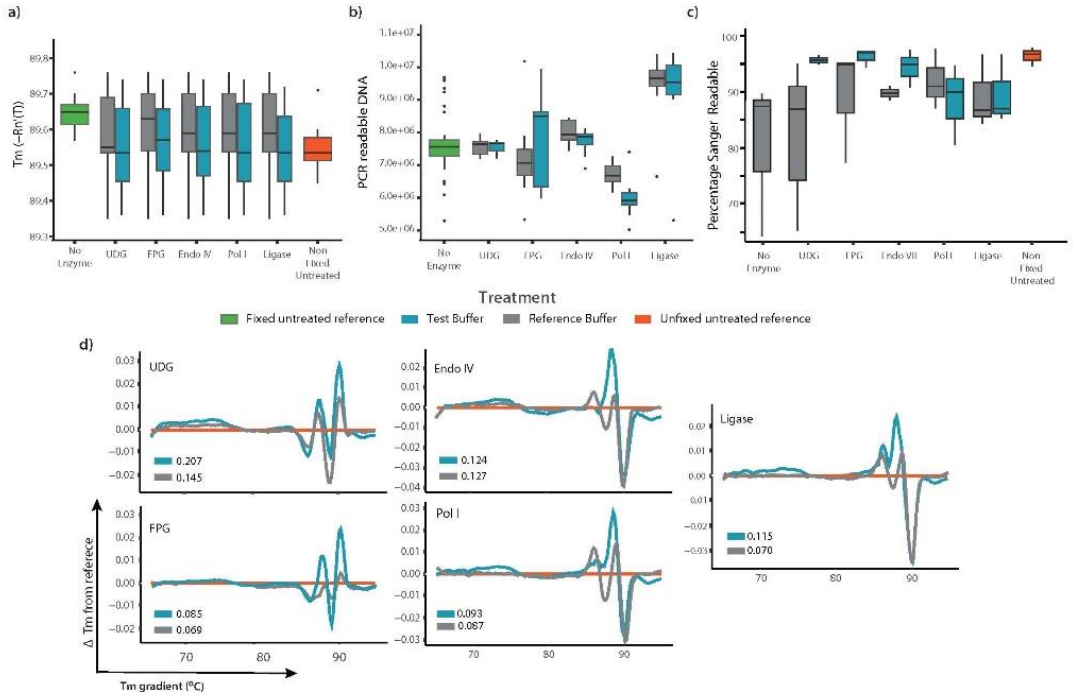
To conclude, the information generated here provides a better understating of FFPE-derived DNA damage, informing strategies for its repair. Here is also presented a thoroughly characterised method to address this damage. Given the increased activity in, and controversy surrounding, the field of low-biomass microbiome analysis, methods that improve the quality of microbiome studies (through sensitivity improvement or access to increased sample size) such as described here, are necessary. Given the paucity of published information on mammalian FFPE DNA repair, and none on bacterial repair, the strategy devised here provides compelling evidence to further pursue BER strategies to improve the sequencing quality of bacterial FFPE DNA and possibly mammalian FFPE DNA.

SUPPLEMENTARY MATERIAL



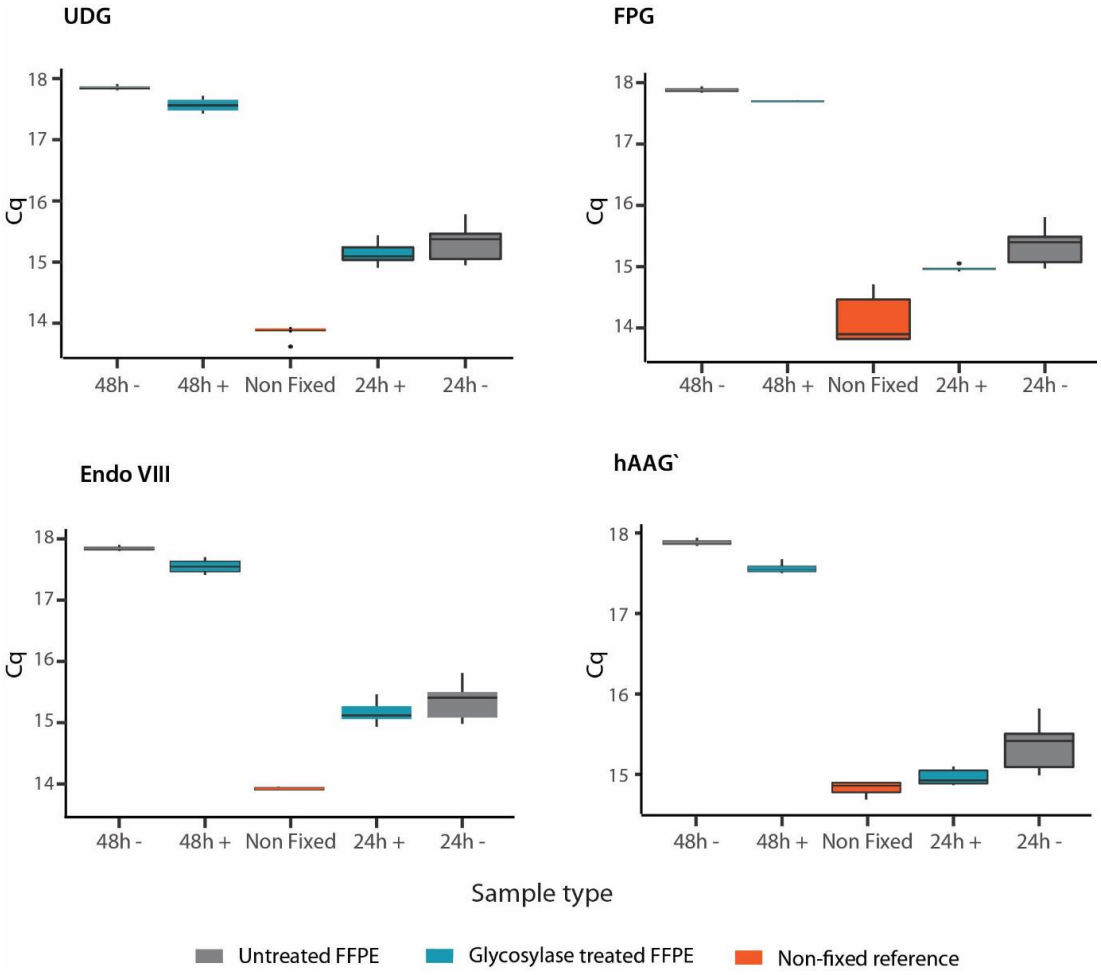
Supplementary Figure 1. Evaluating the effect of DNA glycosylases on bacterial FFPE DNA.

DNA purified from FFPE blocks loaded with *E. coli fixed for 24h or 48h was pooled and equal quantities subjected to treatment with DNA glycosylases shown in plots. Tm analysis of 4 (≈ 100 bp) DNA sequences was performed on normalised quantities of amplifiable DNA. Shown here are the results for two sequences, wherein the melting temperature of fragments tested is compared between untreated DNA (grey, n = 12), NF DNA (orange, for each box n = 6) and glycosylase treated samples (blue, for each box n =6).*



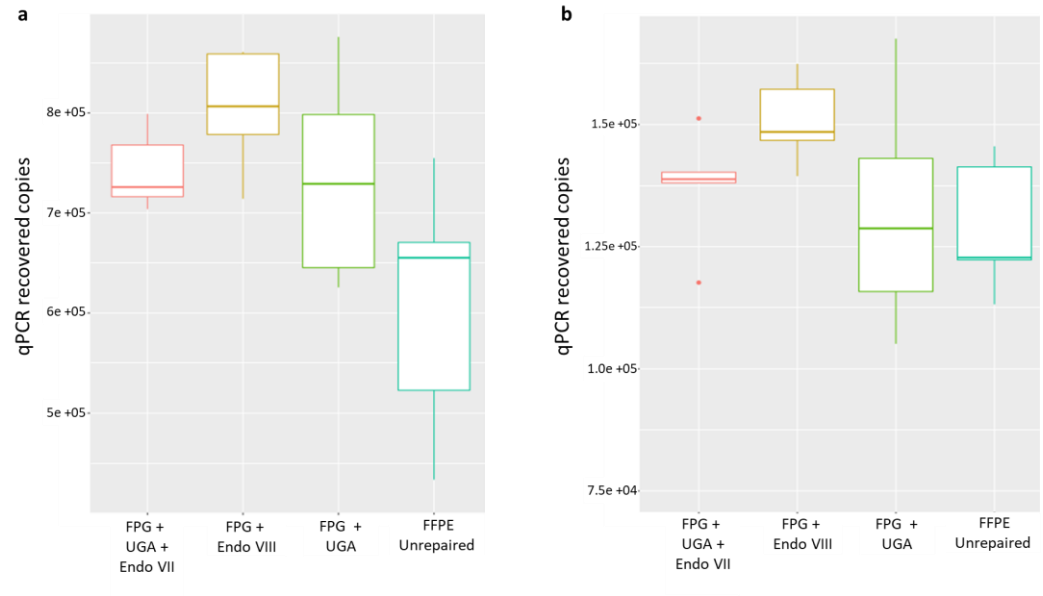
Supplementary Figure 2. DNA repair by BER system: Optimising a buffer.

A universal buffer (blue) allowing the reconstitution of the system was prepared and its influence on enzyme activity assessed by comparing its activity with the buffer provided by supplier(grey). This was analysed by: ***a) Tm analysis*** *(each box n = 6)****, b) Recovery of amplifiable DNA*** *(each box n = 6)****, c) Sanger sequencing readability*** *(each box n = 3)****, d) HRM*** *(each box n = 6)****.*** *In all analysis the outputs of the enzyme activity using both buffers were comparable.*

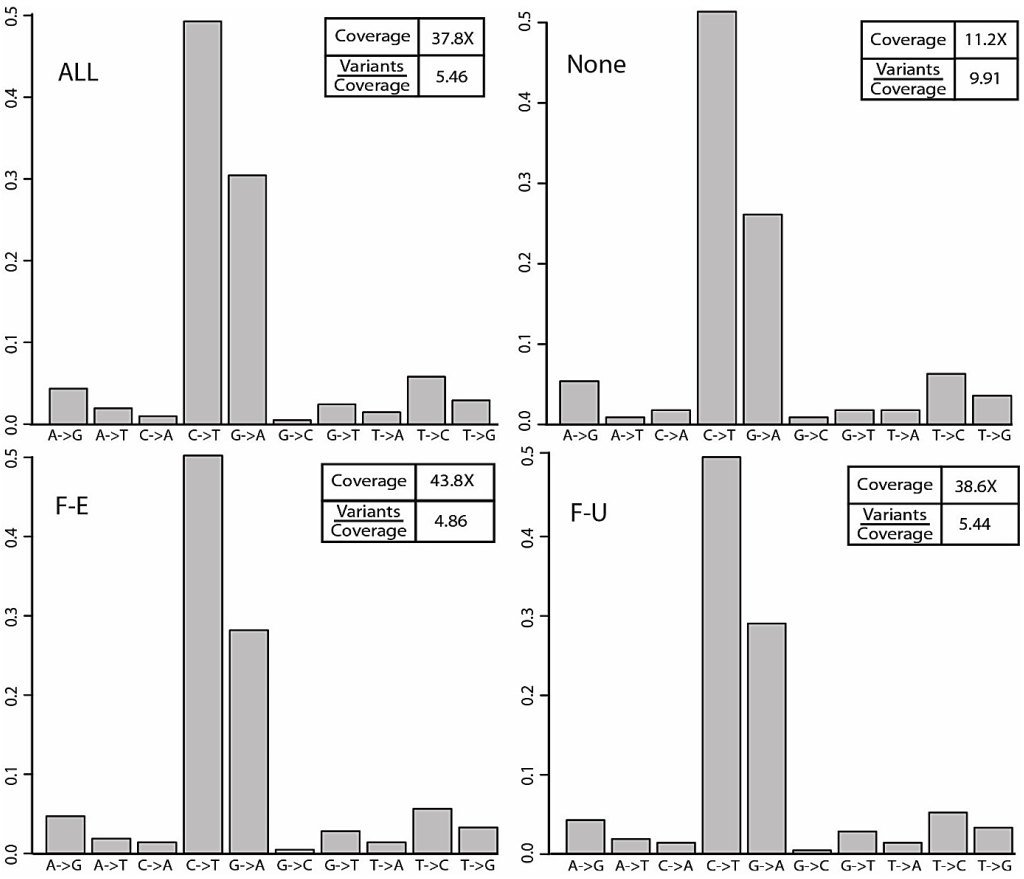


Supplementary Figure 3. Quantitative analysis of treatment with glycosylases.

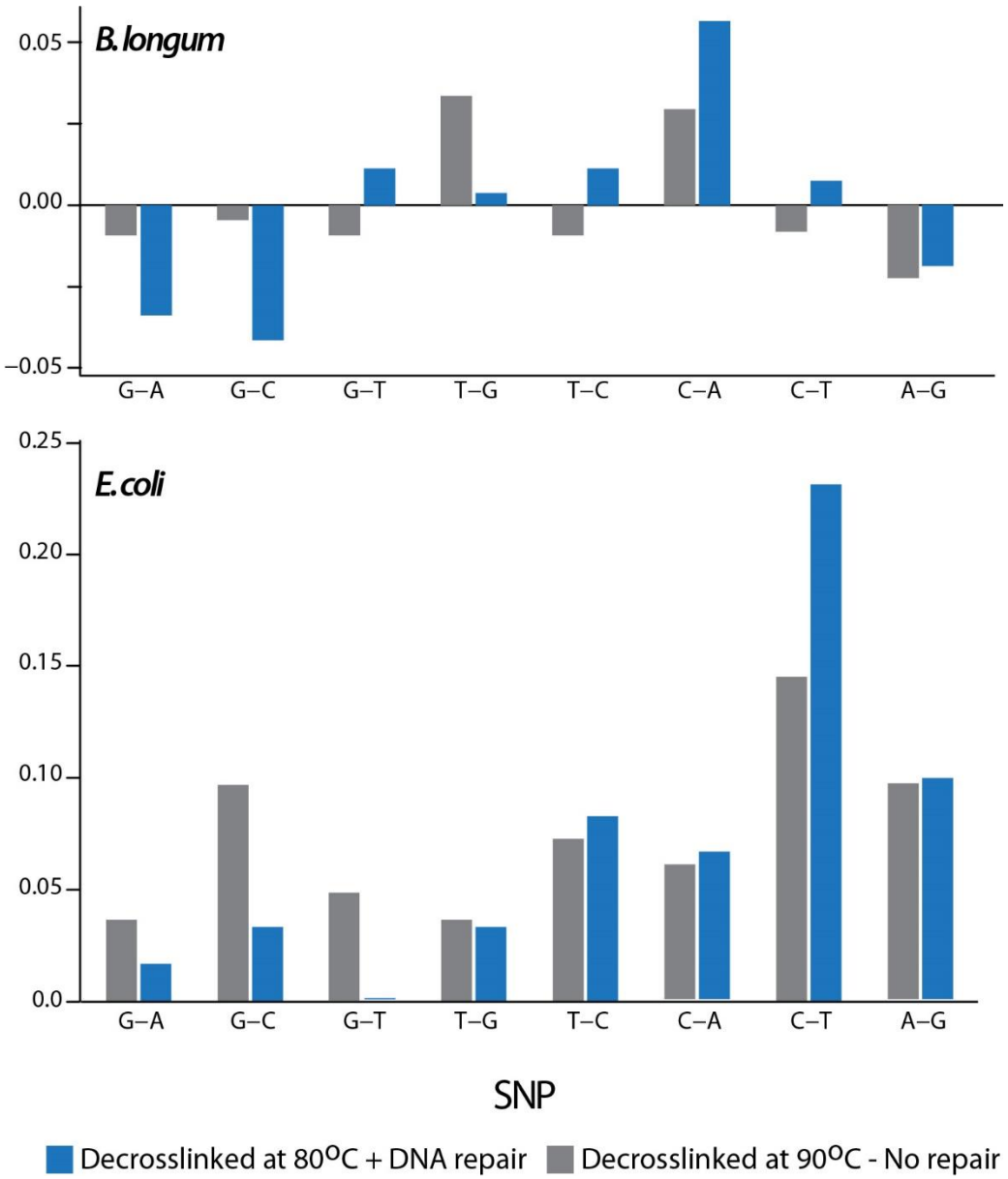
Box plot with average Cq obtained by qPCR after treatment with each glycosylase listed.



Supplementary Figure 4. Quantitative analysis of treatment with single glycosylases BER mixes *(a)* ***Amplification with Taq Polymerase*** *(b)* ***Amplification with Q5.***



Supplementary Figure 5. SNP plots. *Number of variants observed per repair strategy.*



Supplementary Figure 6. SNP variation between E. coli (Gram-) and B. longum (Gram+).

Supplementary Table 1. Specifications of primers used for qPCR assays

Strain/Cell line	Gene/ Accession No	Primer/Probe sequence	F/R*	Product size (bp)
E coli MG1655 [CP032667]	IS5-like element IS5 family transposase AYG17556.1 [CP032667: 230175-231191]	5'TCA TTT GGT CCG CCC GAA AC	F	525
		5'CCA CCA TCA TTG AGG CAC CC	R	
		5'GCC GAA CTG TCG CTT GAT GA	F	217
		5'ATT TGT CTC AGC CGA TGC CG	R	
		5'TCG GCT GAG ACA AAT TGC TC	F	110
		5'GAT GCC AAG AGT GGC CTG	R	
		5'ATG CCA AAG TGC CAC TGA T	F	100
		5'CCA CCA TCA TTG AGG CAC C	R	
		5'CCC CTT GTA TCT GGC TTT CA	F	116
		5'AGA ACA AAA CGG CCA TCA AC	R	
Staphylococcus aureus subsp. *aureus* str. Newman [CP023390.1]	Thermonuclease ATC67584.1 [CP023390.1:1359312-1359845] [72]	5'ACG CCA GAA ACG GTG AAA C	F	533
		5'GAC GTA TTA TTA GCG AAG CCA TAG AGC	R	
		5'CGC CTG TAC AAC CAT TTG GC	F	182
		5'TCT AGC AAG TCC CTT TTC CAC T	R	

*F= Forward primer, R = Reverse primer

Supplementary Table 2. Genomic data from *E. coli* used to calculate DNA glycosylases input.

Genome size	4,636,831 bp
Copy number per ng of DNA	2.102×10^{5}
Moles per ng of DNA	3.49×10^{-19}
Nucleotides per ng of DNA	9.74×10^{11}

Supplementary Table 3. Description of DNA glycosylases tested

Enzyme	Damage targeted	Activity	Product	Units per ng of DNA in reaction	Excised bases	Inactivation
Uracil DNA Glycosylase (Antarctic thermolabile) (UDG)	Deaminated cytosines (dU, 5-OH-dU) [73]	Glycosylase	AP site	0.004	7.20E+10	50°C 5 min
Formamido-pyrimidine DNA glycosylase (FPG)	Oxidised purines (8-oxo-deoxypurines[1], Formamidopyrimidines[2]) Oxidised pyrimidines[3], AP sites [74, 75]	Glycosylase, β, δ - APlyase	5' and 3' P*	0.004	6.03E+09	60°C 10 min
Endonuclease VIII (Endo VIII)	Oxidised Pyrimidines (dT and dU-Glycol, 5,6-dH-dT and dU, 5,6-diOH-dU and dC, 5-OH-6-H-dT and dU, 5-OH-dU and methylhydantoin) Oxidised purine (Fapy-dA) [76]	Glycosylase, β, δ - APlyase	5' and 3' P	0.008	9.13E+10	75°C 10 min
Human Alkyl Adenine DNA Glycosylase (hAAG)	Alkylated purines: 3-me-dA, 7-me-dG, 1,N^6-etheno-dA, hypoxanthine Oxidised purines: deoxy-dI and deoxy-xanthosine [77]	Glycosylase	AP site	0.006	2.97E+10	65°C 10 min

***** *P = phosphates; dT: deoxy-thymine; dA: deoxy-adenine; dC: deoxy-cytosine; dU: deoxy-uracil; dI: deoxy-Inosine; OH: Hydroxy; diOH: dihydroxyl me: methyl, dH: dihydroxyl*

[1] *8-oxodeoxypurines: 8-oxo-dG, 8-oxo-dA, 8-oxo-dNebularine, and 8-oxo-dInosine*

[2]*Formamidopyrimidines: fapy-dG, fapy-dA, and me-fapy-dG*

[3]*Oxidised pyrimidines: 5-hydroxy-deoxycytosine and 5-hydroxy-deoxyuridine.*

Supplementary Table 4. Description of downstream lesion repair enzymes.

Enzyme	For Glycosylase	Activity	Product	Units per ng of DNA in reaction	Repaired bases/ ends	Cofactor
Endonuclease IV (Endo IV)	UDG, hAAG	Removes AP sites [41, 78]	3'OH and 5'dRP*	0.01	4.52E+09	-
T4 Polynucleotide DNA Kinase (PNK)	FPG, Endo VIII	Removes 3' Phosphates [79]	3'OH and 5'P*	0.017	1.25E+10	DTT (5 mM)
DNA Polymerase I (Pol I)	All	3'-5' Exonuclease removes 5'dRP and fills nicks [41, 78]	Nick translation & nucleotide incorporation	0.015	9.74E+16	dNTPs (33 μM)
E coli DNA ligase	All	Gap sealing [41, 78]	PD bond between 5´P and 3'OH	0.025	7.23E+09	NAD+ (50 μM)

***** P = phosphates, dRP = deoxyribose phosphate, PD = phosphodiester bond

REFERENCES

1. Gaffney EF, Riegman PH, Grizzle WE, Watson PH: **Factors that drive the increasing use of FFPE tissue in basic and translational cancer research.** *Biotechnic & Histochemistry* 2018, **93:**373-386.

2. van Beers EH, Joosse SA, Ligtenberg MJ, Fles R, Hogervorst FBL, Verhoef S, Nederlof PM: **A multiplex PCR predictor for aCGH success of FFPE samples.** *British journal of cancer* 2006, **94:**333-337.

3. Blow N: **Tissue issues.** *Nature* 2007, **448:**959-960.

4. Huffnagle GB, Dickson RP, Lukacs NW: **The respiratory tract microbiome and lung inflammation: a two-way street.** *Mucosal Immunology* 2017, **10:**299-306.

5. Marsh RL, Kaestli M, Chang AB, Binks MJ, Pope CE, Hoffman LR, Smith-Vaughan HC: **The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx.** *Microbiome* 2016, **4:**37.

6. Castillo DJ, Rifkin RF, Cowan DA, Potgieter M: **The Healthy Human Blood Microbiome: Fact or Fiction?** *Frontiers in Cellular and Infection Microbiology* 2019, **9**.

7. Stinson LF, Boyce MC, Payne MS, Keelan JA: **The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth.** *Frontiers in Microbiology* 2019, **10**.

8. Ozkan J, Coroneo M, Willcox M, Wemheuer B, Thomas T: **Identification and Visualization of a Distinct Microbiome in Ocular Surface Conjunctival Tissue.** *Investigative Ophthalmology & Visual Science* 2018, **59:**4268-4276.

9. Zhou B, Sun C, Huang J, Xia M, Guo E, Li N, Lu H, Shan W, Wu Y, Li Y, et al: **The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients.** *Scientific Reports* 2019, **9:**1691.

10. Chen J, Douglass J, Prasath V, Neace M, Atrchian S, Manjili MH, Shokouhi S, Habibi M: **The microbiome and breast cancer: a review.** *Breast Cancer Res Treat* 2019.

11. Beck JM, Young VB, Huffnagle GB: **The microbiome of the lung.** *Translational research : the journal of laboratory and clinical medicine* 2012, **160:**258-266.

12. Do H, Dobrovic A: **Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization.** *Clinical Chemistry* 2015, **61:**64-71.

13. Kennedy-Darling J, Smith LM: **Measuring the Formaldehyde Protein–DNA Cross-Link Reversal Rate.** *Analytical Chemistry* 2014, **86:**5678-5681.

14. Einaga N, Yoshida A, Noda H, Suemitsu M, Nakayama Y, Sakurada A, Kawaji Y, Yamaguchi H, Sasaki Y, Tokino T, Esumi M: **Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.** *PloS one* 2017, **12:**e0176280-e0176280.

15. Bonnet J, Colotte M, Coudy D, Couallier V, Portier J, Morin B, Tuffet S: **Chain and conformation stability of solid-state DNA: implications for room temperature storage.** *Nucleic acids research* 2010, **38:**1531-1546.

16. Rait VK, Zhang Q, Fabris D, Mason JT, O'Leary TJ: **Conversions of formaldehyde-modified 2'-deoxyadenosine 5'-monophosphate in conditions modeling formalin-fixed tissue dehydration.** *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* 2006, **54:**301-310.

17. Yoshioka S, Aso Y: **Correlations between molecular mobility and chemical stability during storage of amorphous pharmaceuticals.** *J Pharm Sci* 2007, **96:**960-981.

18. Lindahl T: **Instability and decay of the primary structure of DNA.** *Nature* 1993, **362:**709-715.

19. Vesnaver G, Chang CN, Eisenberg M, Grollman AP, Breslauer KJ: **Influence of abasic and anucleosidic sites on the stability, conformation, and melting behavior of a DNA duplex: correlations of thermodynamic and structural data.** *Proceedings of the National Academy of Sciences* 1989, **86:**3614-3618.

20. Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, Kanapin A, Samsonova A, Vavoulis DV, Ross MT, et al: **Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.** *Genet Med* 2018, **20:**1196-1205.

21. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, Trinh E, MacLeod T, McDonald H, Bala M, et al: **Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples.** *Nucleic acids research* 2019, **47:**e12-e12.

22. Watanabe M, Hashida S, Yamamoto H, Matsubara T, Ohtsuka T, Suzawa K, Maki Y, Soh J, Asano H, Tsukuda K, et al: **Estimation of age-related DNA degradation from formalin-fixed and paraffin-embedded tissue according to the extraction methods.** *Experimental and therapeutic medicine* 2017, **14:**2683-2688.

23. Srinivasan M, Sedmak D, Jewell S: **Effect of fixatives and tissue processing on the content and integrity of nucleic acids.** *The American journal of pathology* 2002, **161:**1961-1971.

24. Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, Kanapin A, Samsonova A, Vavoulis DV, Ross MT, et al: **Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.** *Genetics in Medicine* 2018, **20:**1196-1205.

25. Vitosevic K, Todorovic M, Varljen T, Slovic Z, Matic S, Todorovic D: **Effect of formalin fixation on pcr amplification of DNA isolated from healthy autopsy tissues.** *Acta Histochem* 2018, **120:**780-788.

26. Wong SQ, Li J, Tan AYC, Vedururu R, Pang J-MB, Do H, Ellul J, Doig K, Bell A, McArthur GA, et al: **Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing.** *BMC Medical Genomics* 2014, **7:**23.

27. Munchel S, Hoang Y, Zhao Y, Cottrell J, Klotzle B, Godwin AK, Koestler D, Beyerlein P, Fan J-B, Bibikova M, Chien J: **Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics.** *Oncotarget* 2015, **6:**25943-25961.

28. Oh E, Choi Y-L, Kwon MJ, Kim RN, Kim YJ, Song J-Y, Jung KS, Shin YK: **Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples.** *PLOS ONE* 2015, **10:**e0144162.

29. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ: **Comparison of Clinical Targeted Next-Generation Sequence Data from Formalin-Fixed and Fresh-Frozen Tissue Specimens.** *The Journal of Molecular Diagnostics* 2013, **15:**623-633.

30. Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS, Zemskaya TI: **The effect of 16S rRNA region choice on bacterial community metabarcoding results.** *Scientific Data* 2019, **6:**190007.

31. Lindahl T, Nyberg B: **Rate of depurination of native deoxyribonucleic acid.** *Biochemistry* 1972, **11:**3610-3618.

32. Thanbichler M, Viollier PH, Shapiro L: **The structure and function of the bacterial chromosome.** *Current Opinion in Genetics & Development* 2005, **15:**153-162.

33. Webb CD, Teleman A, Gordon S, Straight A, Belmont A, Lin DC-H, Grossman AD, Wright A, Losick R: **Bipolar Localization of the Replication Origin Regions of Chromosomes in Vegetative and Sporulating Cells of B. subtilis.** *Cell* 1997, **88:**667-674.

34. Selway CA, Eisenhofer R, Weyrich LS: **Microbiome applications for pathology: challenges of low microbial biomass samples during diagnostic testing.** *The Journal of Pathology: Clinical Research*, **n/a**.

35. Wallace SS: **Base excision repair: a critical player in many games.** *DNA repair* 2014, **19:**14-26.

36. Krokan HE, Bjørås M: **Base Excision Repair.** *Cold Spring Harbor Perspectives in Biology* 2013, **5**.

37. Do H, Dobrovic A: **Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase.** *Oncotarget* 2012, **3:**546-558.

38. Hosein AN, Song S, McCart Reed AE, Jayanthan J, Reid LE, Kutasovic JR, Cummings MC, Waddell N, Lakhani SR, Chenevix-Trench G, Simpson PT: **Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP–CGH analysis.** *Laboratory Investigation* 2013, **93:**701.

39. Dalhus B, Laerdahl JK, Backe PH, Bjørås M: **DNA base repair – recognition and initiation of catalysis.** *FEMS Microbiology Reviews* 2009, **33:**1044-1078.

40. Jacobs AL, Schär P: **DNA glycosylases: in DNA repair and beyond.** *Chromosoma* 2012, **121:**1-20.

41. Krwawicz J, Arczewska KD, Speina E, Maciejewska A, Grzesiuk E: **Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease.** *Acta Biochim Pol* 2007, **54:**413-434.

42. Dizdaroglu M, Coskun E, Jaruga P: **Repair of oxidatively induced DNA damage by DNA glycosylases: Mechanisms of action, substrate specificities and excision kinetics.** *Mutation Research/Reviews in Mutation Research* 2017, **771:**99-127.

43. Hegde ML, Izumi T, Mitra S: **Oxidized base damage and single-strand break repair in mammalian genomes: role of disordered regions and posttranslational modifications in early enzymes.** *Progress in molecular biology and translational science* 2012, **110:**123-153.

44. Cronin M, Akin AR, Collins SA, Meganck J, Kim JB, Baban CK, Joyce SA, van Dam GM, Zhang N, van Sinderen D, et al: **High resolution in vivo bioluminescent imaging for the study of bacterial tumour targeting.** *PLoS One* 2012, **7:**e30940.

45. Hughes S, Lau J: **A technique for fast and accurate measurement of hand volumes using Archimedes' principle.** *Australasian Physics & Engineering Sciences in Medicine* 2008, **31:**56.

46. Do H, Dobrovic A: **Limited copy number-high resolution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies.** *Mol Cancer* 2009, **8:**82.

47. Metz B, Kersten GFA, Hoogerhout P, Brugghe HF, Timmermans HAM, de Jong A, Meiring H, Hove Jt, Hennink WE, Crommelin DJA, Jiskoot W: **Identification of Formaldehyde-induced Modifications in Proteins: REACTIONS WITH MODEL PEPTIDES.** *Journal of Biological Chemistry* 2004, **279:**6235-6243.

48. Lu K, Ye W, Zhou L, Collins LB, Chen X, Gold A, Ball LM, Swenberg JA: **Structural characterization of formaldehyde-induced cross-links between amino acids and deoxynucleosides and their oligomers.** *J Am Chem Soc* 2010, **132:**3388-3399.

49. Yudkina AV, Dvornikova AP, Zharkov DO: **Variable termination sites of DNA polymerases encountering a DNA–protein cross-link.** *PLOS ONE* 2018, **13:**e0198480.

50. Kawashima Y, Kodera Y, Singh A, Matsumoto M, Matsumoto H: **Efficient extraction of proteins from formalin-fixed paraffin-embedded tissues requires higher concentration of tris(hydroxymethyl)aminomethane.** *Clinical proteomics* 2014, **11:**4-4.

51. Hoffman EA, Frey BL, Smith LM, Auble DT: **Formaldehyde crosslinking: a tool for the study of chromatin complexes.** *J Biol Chem* 2015, **290:**26404-26411.

52. Kamps JJAG, Hopkinson RJ, Schofield CJ, Claridge TDW: **How formaldehyde reacts with amino acids.** *Communications Chemistry* 2019, **2:**126.

53. d'Abbadie M, Hofreiter M, Vaisman A, Loakes D, Gasparutto D, Cadet J, Woodgate R, Pääbo S, Holliger P: **Molecular breeding of polymerases for amplification of ancient DNA.** *Nature biotechnology* 2007, **25:**939-943.

54. Feuillie C, Merheb MM, Gillet B, Montagnac G, Daniel I, Hänni C: **Detection of DNA Sequences Refractory to PCR Amplification Using a Biophysical SERRS Assay (Surface Enhanced Resonant Raman Spectroscopy).** *PLOS ONE* 2014, **9:**e114148.

55. Jiang X, Jiang X, Feng S, Tian R, Ye M, Zou H: **Development of Efficient Protein Extraction Methods for Shotgun Proteome Analysis of Formalin-Fixed Tissues.** *Journal of Proteome Research* 2007, **6:**1038-1047.

56. Sawyer WH, Puckridge J: **The dissociation of proteins by chaotropic salts.** *J Biol Chem* 1973, **248:**8429-8433.

57. Murphy CL, Eulitz M, Hrncic R, Sletten K, Westermark P, Williams T, Macy SD, Wooliver C, Wall J, Weiss DT, Solomon A: **Chemical Typing of**

Amyloid Protein Contained in Formalin-Fixed Paraffin-Embedded Biopsy Specimens. *American Journal of Clinical Pathology* 2001, **116:**135-142.

58. Chein Y-H, Davidson N: **RNA:DNA hybrids are more stable than DNA:DNA duplexes in concentrated perchlorate and trichloroacetate solutions.** *Nucleic Acids Research* 1978, **5:**1627-1637.

59. Lambert D, Draper DE: **Denaturation of RNA secondary and tertiary structure by urea: simple unfolded state models and free energy parameters account for measured m-values.** *Biochemistry* 2012, **51:**9014-9026.

60. Lu K, Craft S, Nakamura J, Moeller BC, Swenberg JA: **Use of LC-MS/MS and Stable Isotopes to Differentiate Hydroxymethyl and Methyl DNA Adducts from Formaldehyde and Nitrosodimethylamine.** *Chemical Research in Toxicology* 2012, **25:**664-675.

61. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. 4. Mechanism of the initial reaction of formaldehyde with DNA.** *Biochemistry* 1977, **16:**3276-3293.

62. Huang H, Hopkins PB: **DNA interstrand cross-linking by formaldehyde: nucleotide sequence preference and covalent structure of the predominant cross-link formed in synthetic oligonucleotides.** *Journal of the American Chemical Society* 1993, **115:**9402-9408.

63. Feldman MY: **Reactions of nucleic acids and nucleoproteins with formaldehyde.** *Prog Nucleic Acid Res Mol Biol* 1973, **13:**1-49.

64. Marcus Y: **The guanidinium ion.** *The Journal of Chemical Thermodynamics* 2012, **48:**70-74.

65. Chang Y-T, Loew GH: **Reaction Mechanisms of Formaldehyde with Endocyclic Imino Groups of Nucleic Acid Bases.** *Journal of the American Chemical Society* 1994, **116:**3548-3555.

66. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. II. Reaction with endocyclic imino groups of DNA bases.** *Biochemistry* 1975, **14:**1297-1303.

67. Shishodia S, Zhang D, El-Sagheer AH, Brown T, Claridge TDW, Schofield CJ, Hopkinson RJ: **NMR analyses on N-hydroxymethylated nucleobases – implications for formaldehyde toxicity and nucleic acid demethylases.** *Organic & Biomolecular Chemistry* 2018, **16:**4021-4032.

68. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, Zatloukal K, Lehrach H: **Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor**

tissues for copy-number- and mutation-analysis. *PloS one* 2009, **4:**e5548-e5548.

69. Kreutzer DA, Essigmann JM: **Oxidized, deaminated cytosines are a source of C --> T transitions in vivo.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95:**3578-3582.

70. Paabo S, Irwin DM, Wilson AC: **DNA damage promotes jumping between templates during enzymatic amplification.** *J Biol Chem* 1990, **265:**4718-4721.

71. Sikorsky JA, Primerano DA, Fenger TW, Denvir J: **DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency.** *Biochem Biophys Res Commun* 2007, **355:**431-437.

72. Madison BM, Baselski VS: **Rapid identification of Staphylococcus aureus in blood cultures by thermonuclease testing.** *Journal of clinical microbiology* 1983, **18:**722-724.

73. Krokan HE, Drabløs F, Slupphaug G: **Uracil in DNA – occurrence, consequences and repair.** *Oncogene* 2002, **21:**8935-8948.

74. Hatahet Z, Kow YW, Purmal AA, Cunningham RP, Wallace SS: **New substrates for old enzymes. 5-Hydroxy-2'-deoxycytidine and 5-hydroxy-2'-deoxyuridine are substrates for Escherichia coli endonuclease III and formamidopyrimidine DNA N-glycosylase, while 5-hydroxy-2'-deoxyuridine is a substrate for uracil DNA N-glycosylase.** *J Biol Chem* 1994, **269:**18814-18820.

75. Tchou J, Bodepudi V, Shibutani S, Antoshechkin I, Miller J, Grollman AP, Johnson F: **Substrate specificity of Fpg protein. Recognition and cleavage of oxidatively damaged DNA.** *J Biol Chem* 1994, **269:**15318-15324.

76. Dizdaroglu M, Burgess SM, Jaruga P, Hazra TK, Rodriguez H, Lloyd RS: **Substrate Specificity and Excision Kinetics of Escherichia coli Endonuclease VIII (Nei) for Modified Bases in DNA Damaged by Free Radicals.** *Biochemistry* 2001, **40:**12150-12156.

77. Lee C-YI, Delaney JC, Kartalou M, Lingaraju GM, Maor-Shoshani A, Essigmann JM, Samson LD: **Recognition and processing of a new repertoire of DNA substrates by human 3-methyladenine DNA glycosylase (AAG).** *Biochemistry* 2009, **48:**1850-1861.

78. Dianov G, Lindahl T: **Reconstitution of the DNA base excisionrepair pathway.** *Current Biology* 1994, **4:**1069-1076.

79. Dobson CJ, Allinson SL: **The phosphatase activity of mammalian polynucleotide kinase takes precedence over its kinase activity in repair of single strand breaks.** *Nucleic Acids Research* 2006, **34:**2230-2237.

CHAPTER 4:

Development of a novel protocol for bacterial DNA extraction from FFPE samples

ABSTRACT

Background. The role of the microbiome in health status is an expanding research area. Recently, body sites previously considered sterile have been found to harbour an endogenous microbiome. One of the key rate limiting factors in progression of such research is difficulty in accessing sufficient tissue samples for statistically significant analysis to be carried out or to perform retrospective analyses. FFPE tissue represents the biggest repository of human tissue samples and could represent a vital resource for expanding microbiome research. Current methods for isolation of DNA from FFPE samples are not suitable for bacterial microbiome studies as they have been developed for human DNA. As such, we sought to develop a method for processing of FFPE samples to yield a higher quantity and genus range of bacterial DNA than currently available methods, of the quality required for 16S sequencing and whole genome sequencing.

The method consists of: 1) Dewaxing and rehydration with ethanol; 2) Host depletion with Saponin and Benzonase nuclease; 3) Bacterial lysis with Metapolyzyme; 4) Protein digestion with Proteinase K; 5) Decrosslinking with a GuHCL based lysis buffer; 6) a silica column based DNA isolation; 7) DNA repair using the BER pathway. The method was validated using Protoblocks, FFPE murine models, and clinical human tissue samples. DNA quantity and quality in terms of fragment length and sequence fidelity was assessed by qPCR and16S sequencing. The method developed shows clear and significant improvement over the current gold standard in both mock communities and murine samples, this was seen particularly in terms of consistent bacterial lysis across a number of species, and effective host depletion. The tissue dissociation step requires optimisation, and additional measures must be implemented to limit the effect of environmental contamination. Future work to remedy these issues is discussed in the main text. This novel method opens the door for reliable use of standard clinical FFPE tissue samples for modern bacterial sequencing studies.

INTRODUCTION

The use of formalin-fixed, paraffin-embedded (FFPE) tissues in microbial surveys has the potential to revolutionise the field of human microbiome research with unprecedented access to samples. It is well established that a DNA isolation method, prior to sequencing, should be based upon specific study aims, target organisms and sample types [1]. However, at present, no such method exists for bacterial DNA in FFPE samples, although several groups have carried out metabarcoding surveys of bacterial communities within these tissues [2]. There is a plethora confounding features present when carrying out sequence-based analysis of bacterial communities [3], and when coupled with the criticisms levelled at recent sequencing experiments targeting similarly challenging sample types [4] it is unlikely that large scale microbiome studies using FFPE samples will remain tenable without the development of bespoke methodologies and biological standards. The key characteristics of FFPE samples that impair effective microbial analysis are: *(i)* Formalin-derived crosslinks and damages to DNA present in the sample[5]. *(ii)* A high ratio of host to bacterial DNA[6]. *(iii)* DNA extraction methods for FFPE DNA to date are optimised for human cells. *(iv)* The extent of processing necessary leaves samples vulnerable to contamination. *(v)* No standards exist to validate the effects of the above on downstream analysis.

Previous work presented here was has sought to address some of the above issues, namely the design of an effective DNA repair strategy (Chapter 3) and a FFPE-based biological standard to validate the effectiveness of any developed methodology (Chapter 2). This study draws on these tools, and combines them with effective and validated host depletion and bacterial lysis strategies, to present a final method for the analysis of bacterial communities within FFPE tissues.

Host Cell Depletion. In low bacterial biomass samples, such as many human tissue scenarios, host DNA can constitute > 90 % of total DNA, severely limiting metagenomic studies, as the vast majority of sequencing reads are taken up by this background human DNA. This is critical, particularly for whole genome shotgun (WGS) methods [7]. It has been also shown to affect the outputs of 16S rRNA amplicon sequencing, since in reactions of low bacterial to human DNA ratios, human

DNA can be annealed and amplified during 16S PCR [8]. Furthermore, a reduction in bacterial range and rare bacteria taxa can occur during dilutions made to avoid overloading DNA in PCR reactions [9]. Therefore, any reduction in the ratio of background mammalian to target bacterial DNA would improve readout. DNAse treatment can reduce the quantity of intact background DNA, if it's activity can be targeted to mammalian cells, e.g. by restricting access of the DNAse enzyme to only mammalian cells. Mammalian specific-membrane permeabilisation may achieve this.

For differential membrane permeabilisation, structural differences between mammalian and bacterial cells can guide the choice of permeabilisation agents [10]. In principle, both mammalian membranes and the outer membrane (OM) of Gram-negative (G-) bacteria are composed of a phospholipid bilayer [11, 12]. However, in the OM of G- bacteria, the phospholipid bilayer is surrounded by an outer envelope of lipopolysaccharide (LPS) with polysaccharide chains facing the hydrophilic end [13]. LPS are densely and tightly packed highly hydrophobic structures that seal the inner-membrane from action of detergents [14]. The lipid bilayer of mammalian cells is made mainly of phospholipids, with variable contents of glycosphingolipids and cholesterol rafts [11, 15].

Several host depletion strategies for microbiome analysis have been published [16]. However, the principles by which mammalian membranes are lysed in these cases, do not apply for FFPE samples, where dead cells have no membrane potential or active homeostatic mechanisms to ensure tonicity. FFPE tissue is also hardened by the nature of the processing, such that methods that lyse membranes with soft-tissue lysis beads are not suitable for this sample type [17]. Chaotrophic agents are capable of disrupting hydrophobic interactions [18, 19], such as those maintaining the tightly packed LPS structure in G- bacteria, annulling the protection of the phospholipid bilayer [14], as in the case of mammalian membranes [20], exposing intermembrane proteins that are easily denatured by these agents, and the peptidoglycan layer [21]. The state of the interpeptide bridges enforcing the structure of the peptidoglycan layer could be severed by formalin fixation [22]. Finally, methods developed to capture host DNA by binding CpG islands in mammalian DNA [18, 23] were deemed unsuitable for FFPE DNA, for which CpG sites represent hotspots for sequence alterations, with higher number of degraded cytosines than in NF samples observed in multiple studies.

Non-ionic, mild detergents, are known to solubilise membrane lipids, without significantly disrupting the integrity membrane proteins [10]. Among these, Triton-X and Tween-20 preferentially solubilise membrane phospholipids [24]. Alternatively, Saponin and Digitonin target cholesterol, present in high ratios in the mammalian cellular membrane [25]. It has been shown that these detergents have virtually no activity in solubilising membranes without cholesterol (nuclear envelope, vacuoles and mitochondria) [25, 26], making them selective for mammalian cell membranes. All of these detergents have been shown to temporarily permeabilise membranes, inducing pores 4 – 5 nm of diameter in live cells at low concentrations and completely solubilising membranes it at higher concentrations [24, 27]. The efficacy of these detergents upon FF mammalian cells is well described and has been found variable [28]. While live bacteria are tolerant to these detergents [29, 30] their effect on FF bacteria is still unexplored.

Bacterial Lysis. Bacterial lysis is a critical step in sample processing for microbiome analysis. It can be major source of bias in community composition, as lysis methods that favour particular taxa will cause overrepresentation in the final analysis [8, 31-35]. Many methods for unbiased bacterial lysis of non-fixed samples have been proposed and applied, including bead-beating, enzymatic lysis, detergents and denaturing agents [32, 33, 36]. Recently, several studies have agreed that bead-beating is the lysis method that yields higher uniformity of bacterial lysis and have shown that combining bead-beating with other methods shows further improvements in uniformity [36, 37]. Furthermore, it has been found that properties inherent to the type of sample influence the efficiency of the sample prep [34].

FFPE samples are characterised by DNA damage that includes high levels of fragmentation and DNA damage reducing the recovery of PCR/sequencing readable DNA [38, 39]. Additionally, FFPE samples typically have low bacterial biomass concealed by large quantities of DNA from the larger human genome. Bead-beating decreases DNA yields by causing DNA fragmentation leading to the formation of chimeras during PCR [1, 40, 41], which would be detrimental for FFPE samples. For this sample type, lysis must be performed under conditions that do not negatively affect the integrity of DNA, such as enzymatic lysis. Accordingly, the *Association of Biomolecular Resource Facilities Metagenomics Research Group* developed a mix of

6 lytic enzymes (achromopeptidase, chitinase, lyticase, lysostaphin, lysozyme, and mutanolysin) that target the cell wall of bacteria, yeast, and fungi, and is able to lyse recalcitrant endospores [42]. The incorporation of this enzyme, known as Metapolyzyme (Sigma-Aldrich), in sample preparation has been shown to increase the recovery sphearoplasts or protoplasts, and improve the overall DNA recovery across taxa in multiple sample types [1, 42]. Recently, a microbiome study was performed on ancient DNA specimens (with similar levels of DNA damage as FFPE), validating the efficacy of Metapolyzyme over traditional bead beating methods in this sample type [43].

Study Aims. In this study, host DNA depletion, cross-taxa bacterial lysis, were optimised and combined with DNA repair as a single protocol as per Figure 1. Differential cell lysis strategies were investigated and individually validated before a final validation of the method as a whole was performed, using a cancer line and multiple bacterial genera. The combined methodology was assessed using the "Protoblock" biological standard, and by formalin fixed mouse faeces as a high biomass sample. The bespoke protocol was compared with the current gold standard, the Qiagen QIAmp FFPE kit. Lastly, low biomass samples of malignant formalin fixed patient breast tissue were processed using the novel method, and compared with paired fresh frozen samples.

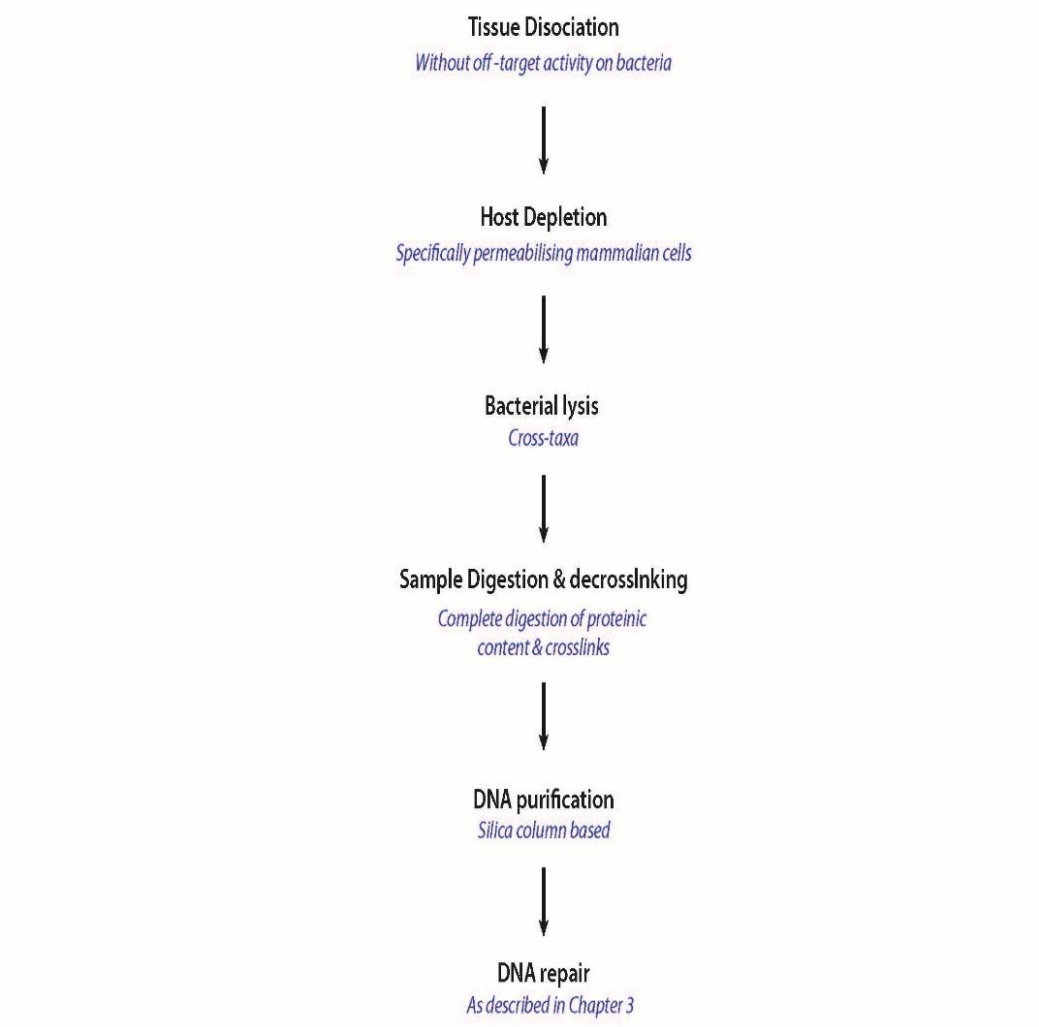


Figure 1. Full Protocol for bacterial DNA isolation from FFPE samples *– describing steps for process and in blue the requirements for the step*.

METHODS

1. Models

The models used to test the different steps of this protocol, were, *a) ex vitro*: *i)* Formalin fixed (FF) cells, *ii)* Protoblocks, and; *b) ex vivo* models: *i)* mice tumours and normal gut tissue and ii) mice faeces.

A. Cellular models

For this type of model, bacterial and mammalian cells were grown, harvested, formalin fixed and counted, as follows:

Cell culture. *Mus musculus* mammary gland cancer cells (4T1) were grown at 37ºC 5% CO2, in RPMI media supplemented with 10% FBS, 100 U/mL penicillin and 100 µg/mL of streptomycin (ThermoFisher) to a final count of 10^8 cells. The cells were harvested with 0.5 ml/10cm^2 trypsin, washed with PBS, pooled and counted with a NucleoCounter® NC-100™ (chemometect, Copenhagen) following manufacturer's instructions. The cells were fixed in 40 ml of 4% buffered formalin for 48h at RT, unless specified.

Bacterial growth conditions. *E. coli* K12 MG1655 carrying a P16Lux plasmid [44] or *E coli* Nissle 1917, was grown aerobically at 37ºC to an OD_{600} of 0.8 in LB medium supplemented with 300 µg/ml Erythromycin. *Staphylococcus aureus* newman (ATCC 25904) was grown aerobically at 37ºC to an OD_{600} of 0.8 in Tod-Hewwit broth. *Bifidobacterium longum* 35624 was anaerobically grown at 37ºC for 24 h in MRS medium. *Lactobacillus amylophilus* (ATCC® 49845™) was grown at 30ºC in an atmosphere of 5% CO2 for 24 h. *Bacteroides thetaiotaomicron* (ATCC®29741™) was grown anaerobically at 37ºC for 24 h in FAB medium (NEOGEN, Lancashire, UK). Bacterial cultures were harvested by centrifugation at 3000 x g, for 10 min at 4ºC, and suspended to 2X with PBS. A one ml aliquot of the suspension was kept for obtaining the viable CFU by retrospectively counting plated dilutions (10^{-5}, 10^{-6}, 10^{-}

^{7}), the rest of the suspension was pelleted and suspended to a 2X concentration with buffered formalin and fixed as specified for each model or experiment.

Counting total fixed bacterial cells. The cell suspension was counted using a bacterial counting kit for flow cytometry (Invitrogen), following manufacturer's instructions. In brief, after fixation, bacterial suspensions were harvested and suspended in 4% Neutral Buffered formalin to a suspension of an approximate density of 10^6 cells/µl (100X concentration). The final suspension volume was measured, to account for the displacement created by the cells in the solution. Displacement volume was calculated by subtracting the volume of formalin added from the final suspension volume. This was 4 x 10^6 µl/cell for 4T1 cells, 1.04 x 10^7 µl/cell for *E. coli,* 1.38 x 10^7 µl/cell for *S. aureus*, 1.26 x 10^6 µl/cell for *B. longum*, 4.32 x 10^7 µl/cell for *B. thetaiotaomicron* and 1.37 x 10^7 µl/cell for *L. amylophilus*. A 10% aliquot was taken from this suspension and serially diluted (100X) with filtered sterilised 0.15M NaCl solution to obtain a cell density of approximately 1x 10^6 cells in 989 µl of NaCl. Bacterial cells were stained with 1 µl of SytoBC and 10 µl (1X 10^6) of counting beads were added to the suspension. Cells were counted in an LSR II Flow Cytometer (BD Biosciences, NJ, USA). The acquisition trigger was set to side scatter and regulated for each bacterial strain to filter out electronic noise without missing bacterial cells. This value was approximately 800.

Protoblocks. Protoblocks with the bacterial and 4T1 cell content specified per experiment were made following the same protocols described in Chapter 2. Briefly, cells formalin fixed in formalin for 18 h were suspended to a density of 2 – 6 x 10^6 bacterial cells per µl and 1.2 x 10^5 4T1 cells per µl. The volume corresponding to 2 x 10^7 CFU for each bacterial strain and 2.2 x 10^7 4T1 cells were aliquoted to create a mixed cell suspension. These suspension was mixed thoroughly with an equal volume of sterile agar (1.5X concentration) and the mix pipetted to a cylindrical mould made from a 54 x 11 mm adapter tube (SARSTEDT, Cat No. 55.1570), with a flat end sealed with a double layer of parafilm, and let solidify for 3 minutes. The parafilm was removed from the bottom and the disk shaped cell matrix dropped into a 15 ml tube filled with 8 ml of formalin, using a sterile bacterial loop. The protoblocks were fixed for an extra 24 h for 48 h fixation time point or processed immediately for 24h fixation time point. After fixation, the cassettes containing the guts were dehydrated and

paraffin embedded automatically with a LOGOS J (Milestone Medical, Bergamo). Following this protocol: *Dehydration* for 4 hours with increasing concentrations of ethanol at 37ºC. *Clearing* with 2 x xylene washes for 2 h and 20 min each, at 37 oC, 2 x washes of isopropanol for 1 h and 40 min at 37ºC and 1 x wash with isopropanol for 50 min at 60 ºC. P*araffin embedding* for 8 h and 32 min at 62 ºC. Once dehydrated and paraffinised, the protoblocks' volume, diameter and height were measured with a calliper and by volume displacement.

Sectioning. The blocks were sectioned keeping an aseptic technique either at 4 µm for imaging or at 10-20 µm for DNA purification. The cell load of each slide was calculated by multiplying the total bacterial load by the volume of each slide, using the volume of a cylinder.

Microscopy. Following protocols from Chapter 2, cell integrity was evaluate by H&E staining with Mayer's haematoxylin (Sigma, MHS16) and Gram-staining (Sigma, 77730). To confirm the bacterial content of each protoblock sections and three immunofluorescent stained with DAPI, *α-S. aureus* or *α- E. coli* sections were used to label bacteria and 4T1 cells. 25 fields of view were counted for each slide and the average plotted against the slide volume to obtain cell density per µl of block.

B. Murine models

Mice. BALB/c mice were housed in a conventional environment (temperature 21 ℃, 12 h light: 12 h darkness, humidity 50%). They were fed a standard non-sterile pellet diet and tap water ad libitum. Mice were allowed 2 weeks to acclimatise before entering the study. All animal procedures were performed according to national ethical guidelines following approval by the University College Cork Animal Experimentation Ethics Committee.

Mice gut tissue processing. Distal guts from 2 mice were dissected using an aseptic technique. The gut tissue was opened longitudinally, excess faecal matter removed and the tissue was rolled and placed into a processing cassette, where it was formalin fixed for 24 h. After fixation, the cassettes containing the guts were dehydrated and paraffin embedded using the same protocol described for protoblocks. Processed

tissues were placed in 2 x 2 cm embedding mould and mounted to a processing cassette, using standard histology procedures.

Mice faeces blocks. 6 mice were samples for this model. From each mouse, 3 pellets were collected into a 2 ml Eppendorf tube filled with 1.5 ml of formalin. The tube closed and the pellets fixed for 18 h. After fixation, using a sterile bacterial loop, the 3 pellets were placed in the same cylindrical mould as specified for protoblocks. 350 µl of sterile, molten agar, kept in aliquots at 65ºC was poured onto the pellets and let solidify for 3 minutes. Just as with protoblocks, the disk shaped matrix containing the pellets were dropped into a 15 ml tube filled with 8 ml of formalin. The disks were either processed immediately or further fixed for 24 h (for 48 h fixation) and processed (dehydrated and paraffinised) as specified for protoblocks. The resulting paraffinated disk was placed into 1.5 x 1.5 cm embedding mould and mounted into a cassette using standard histology procedures.

Sectioning. The blocks were sectioned keeping an aseptic technique either at 4 µm for imaging or at 10-20 µm for DNA purification.

Microscopy. The presence of bacterial cells in gut tissue and mouse pellets was evaluated in three Gram-stained sections and the integrity of the tissue cells with H&E.

2. DNA analysis

Conventional PCR. 25 µl reactions were setup using Taq 2X Master Mix (NEB, Ipswich, MA, USA) and 0.25 µM of each primer. Cycling conditions included: an initial denaturation for 30 sec at 95ºC. 25 – 35 cycles of denaturation at 95ºC for 10 sec, annealing for 15 sec at the primers' optimal temperature [54-56ºC] (specified by NEB's calculator for Taq DNA polymerase), 20-40 sec of extension at 68ºC (20 sec for 200bp amplicons and 40 sec for 400-500 bp amplicons), and a 5 min final extension at 68ºC. 10 µl of amplified products were loaded to a 1.5% agarose gel, run at 200V for 20 min, and imaged with Gel Doc EZ System (Bio-Rad)

Quantitative PCR. For dye-based qPCR, reactions were prepared using LUNA Universal qPCR (NEB, USA) and 0.25 µM of each primer and 0.25 µM of probe.

Multiplex qPCR, reactions were prepared using LUNA Universal Probe qPCR (NEB, USA) and 0.5 µM of each primer and 0.25 µM of probe for each strain. Reactions for simultaneously quantifying three bacterial strains, were set using fluorochromes: FAM, HEX, CY3. Amplification was performed in an AriaMx (Agilent Technologies, USA) using fluorescence probe or DNA binding dye absolute quantitation experiment type.

The thermal profile included an initial denaturation of 1 min at 95ºC, and 40 cycles of denaturation at 95ºC for 10 sec, annealing for 15 sec at the primers' optimal temperature [54-56ºC] (specified by NEB's calculator for Hot Start Taq) and 20-40 sec of extension at 68 ºC (20 sec for 200bp amplicons and 40 sec for 400-500 bp amplicons).

Each assay included triplicates of 5 points standards using log-dilutions of gene blocks (750 bp), which were designed based upon species-specific genetic regions. Primers and/or probes targeting these regions and maintaining a similar Tm (+/-2ºC) were designed using the NCBI primer design tool and their parameters (ΔG, hairpins and dimers) verified using IDT's Oligo analyser tool. Primers and gene-blocks were acquired from IDT (Coralville, USA) (see Table 1). qPCR efficiencies between 95% and 105% and R-square values higher than 0.995 were deemed as acceptable, all samples were ran in triplicate.

Table 1. Specifications for primers and probes utilised

Strain/Cell line	Gene/ Accession No	Primer/Probe sequence	F/ R/	Product size (bp)
E coli MG1655 [CP032667]	IS5-like element IS5 family transposase AYG17556.1 [CP032667: 230175-231191]	5'TCA TTT GGT CCG CCC GAA AC	F	525
		5'CCA CCA TCA TTG AGG CAC CC	R	
		5'GCC GAA CTG TCG CTT GAT GA	F	217
		5'ATT TGT CTC AGC CGA TGC CG	R	
E coli Nissle 1917 [CP022687]	plasmid pNissle1 [CP022687] [45]	5'GAA CAT ACA GAC CGC TAT CC	F	460
		5'GCC TCT GTA AGC TCT CTA ATG	R	
		56-FAM/CTTGATGAC/ZEN/CTGACGATGTTGAGC/3IABkFQ/	P	
Staphylococcus aureus subsp. *aureus* str. Newman [CP023390.1]	Thermonuclease ATC67584.1 [CP023390.1:1359312-1359845] [46]	5'CGC CTG TAC AAC CATT TGG C	F	182
		5'TCT AGC AAG TCC CTT TTC CAC T	R	
		5'TGC TAT GAT TGT GGT AGC CAT C	F	425
		5'ACT TCT CTC TAG CAA GTC CCT	R	
		5'Cy3/CAA GAT CGC TAT GGT AGA ACA TTG GCG TAT G/3BHQ_2/	P	
Lactobacillus amylophilus (ATCC® 49845™) [1423721]	hsp60 gene [HE573891.1: 314172 - 314752] [47]	5'CCC TTG GAA CGT GGT TAT G	F	474
		5'ACG GGT TCC TTC GACT T	R	
		5HEX/CCA TTA AAC/ZEN/AGG GTC GTG GGT ATG/3IABkFQ/	P	
Bacteroides thetaiotaomicron (ATCC®29741™) NZ_CP012937.1	NP_809948.1 [NC_004663.1: 1306764-1307540] [48]	5'CAT TCT TCT TCT TGT GGC TAA AC	F	480
		5'TGG GAA ATG TAC AAC CTG AAA	R	
		56-FAM/TGA GCT TGG/ZEN/GCT ATT TGC TGT TTA/3IABkFQ	P	
Bifidobacterium longum strain 35624	Glycosyl transferase CP013673 [461108-461924][49]	5'GTC GGA CTT GCT GCG TTT ATC GTT G	F	125
		5' CGG GGC GCT TGA TAG AGA ACA ATG	R	
		5'CGT CGT CGT CTG ATT CGT AAG	F	440
		5'GGG CGC TTG ATA GAG AAC AA	R	
		5HEX/CTA TAA GGT /ZEN/AAA TCT TCC AGC CGT ACC GGA G/3IABkFQ/	P	
4T1 cells [ATCC® CRL-2539™] *Mus musculus* [10090]	BetaActin AC144818.4 [NC000071.6: 73696- 73082]	5'GCG TGT CAG ACG TTT TTC CC	F	156
		5' AGA AAA GAG CGG AGG TTC GG	R	
		5'GAT TAC TGC TCT GGC TCC TAG	F	147
		5'GAC TCA TCG TAC TCC TGC TTG	R	
		5'/HEX/CTG GCC TCA/ZEN/CTG TCC ACC TTC C/3IABkFQ/	p	
Bacteria	V3-V4 hypervariable region of 16S rRNA gene	5'TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG	F	460
		5'GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CHV GGG TAT CTA ATCC	R	

16S library Preparation. Genomic DNA was amplified using 16S rRNA gene amplicon polymerase chain reaction (PCR) primers targeting the hypervariable V3-V4 region of the 16S rRNA gene (see Table 1) using the Illumina 16S rRNA gene Sequencing Protocol (Illumina, CA, USA). The amplification reaction was performed in a final volume of 50 µl, containing 1X concentration of NEBNext High Fidelity 2X PCR Master Mix (NEB, USA), 0.5 µM of each primer, 8 µl template (5-15 ng/µl) and 12 µl nuclease free water.

Thermal cycling was performed in an Eppendorf Mastercycler, with a thermal profile that included a 98 ^{o}C denaturation for 30 sec and 25 cycles of 98 ^{o}C for 10 sec, annealing at 55 ^{o}C for 30 sec and extension at 72 ^{o}C for 30 sec. A final extension step was performed at 72 ^{o}C for 5 min. Amplification was confirmed by running 5 µl PCR product on a 1.5 % agarose followed by imaging on a Gel Doc EZ System. The product was approximately 450 base pairs (bp) in size.

PCR-positive products were cleaned per the 'PCR Clean-Up' section of the Illumina protocol. Sequencing libraries were then prepared using the Nextera XT Index Kit (Illumina) and cleaned per the Illumina protocol. Libraries were quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay, normalised to 10 ng/µl and pooled into a single reaction tube. Further processing was performed by GENEWIZ (Leipzig, Germany) where samples underwent a paired-end 450 bp run on the Illumina MiSeq platform.

3. Host depletion strategy

Experiments to develop a host depletion were performed with flow cytometry and confirmed by qPCR. First, the effect of permeabilisation agents was investigated and later the internalisation and effect of DNAse enzymes.

Membrane permeabilisation assay. For mammalian cells, 8 x 10^7 4T1 cells were fixed in 40 ml of formalin for 48 h, pelleted at 250 x g for 10 min, washed once with TBS (50 mM Tris, 150 mM NaCl, pH 7.6), and suspended to a final density of 2.5 x 10^6 cells per ml in TBS. For bacterial cells, 5 x 10^9 *E. coli* cells, were fixed in 50 ml of formalin for 48 h, pelleted at 300 x g for 10 min, washed once with 20 ml of TBS

and suspended to a final density of 2.5 x 10^7 cell per ml in TBS. 500 µl of the cell suspensions were aliquoted into 1.5 ml tubes and treated with a permeabilisation agent (see Table 2).

The cells were permeabilised for 25 min, at 25ºC, in a thermomixer, shaking at 500 rpm. Permeabilised cells were washed once with TBS and blocked on ice with TBS + 1% BSA. Blocked cells were exposed to 0.75 µg of Cy5 labelled Streptavidin (SAv-Cy5, MW = 60 KDa, globular structure) (Biolegend, CA, USA) for 30 min at 25ºC, shaking at 280 rpm. Cells were washed with 1 ml of 0.15 M NaCl solution and resuspended in 350 µl of the same solution for analysis. Bacterial cells were labelled with 1 µl of SytoBC (Invitrogen) for 5 min and analysed by flow cytometry in a BD LSRII.

4T1 Cells were identified and gated based on their Forward/Side scatter and E. coli cells were detected using the 488-1 (FITC) 525/50 filter for SytoBC and gated using the side scatter. Cy5+ cells were detected with the red 670/14 filter. 10,000 events were recorded for 4T1 cells and 100,000 for bacteria.

Table 2. Permeabilisation agents tested and concentrations (Sigma-Aldrich)

Permeabilisation agent	**Concentration[50]**
Triton X-100	0.1% v/v
Tween-20	0.2% v/v
Saponin	0.1% w/v
Digitonin	0.5 µg/ml

DNAse screening. A screen for selecting the DNAse that had highest activity in depleting DNA in a reaction buffer containing Saponin. DNAses tested: Recombinant DNAse I [1-2U, 1 µl] (Sigma-Aldrich), Turbo DNAse [2U, 1 µl] (Thermo-Fisher), Molysis DNase [2 µl] (Molzym GmbH & Co, Bremen, Germany), RQ1 DNAse [20U, 20 µl] (Promega), Benzonase [75 U, 0.3 µl] (Sigma-Aldrich). 5 x 10^6 4T1 cells, FF for 48h were treated with 0.1% Saponin and the DNAse tested. Reactions were set in

reaction buffers provided or suggested by supplier, supplemented with 0.2% Saponin, for 20 min at 37ºC. The reaction was stopped by either: the addition of EDTA (Benzonase), the supplied reaction Stop Buffer, or by incubating at 75ºC (DNAse I). After which, cells were subject to DNA purification with the QIAGEN DNA mini kit. After which DNA yield was measured by QUBIT. All reactions were performed in triplicate. A no-DNAse control was included with incubated under the same conditions with buffer supplied for DNAse I.

Saponin Titration. Different w/v saponin concentration (0.1%, 0.25%, 0.5%, 1%) were tested in 1 x 10^6 *E. coli cells* that were fixed, washed, permeabilised, blocked and imaged as described for membrane permeabilisation assay.

DNA depletion assay. Cells were fixed, washed and permeabilised as described for the membrane permeabilisation assay. 2.5 x 10^5 4T1 or 2.5 x 10^6 *E. coli* cells were permeabilised, blocked with 500 µl of 1% BSA in TBS+ $MgCl_2$ (20mM Tris-HCL, 20 mM NaCl, 2mM MgCl2, pH 8) for 30 min on ice. Blocked cells were treated with 1.5 µl (≥ 375 units) of Benzonase nuclease (Sigma-Aldrich) for 30 min at 37ºC, shaking at 360 rpm. Treatment was stopped by the addition of 100 µl of 100 mM EDTA. The cells were washed once with TBS and suspended in 0.15M NaCl, where they were stained with 10µM CytoPhase Violet (Biolegend) for 1.5 h at 25ºC, shaking at 200 rpm in the dark. Bacterial cells were labelled with 100 µM of BacLight red (Invitrogen) for 15 min at 25ºC, shaking at 200 rpm and analysed by flow cytometry. 4T1 Cells were identified and gated based on their Forward/Side scatter and E. coli cells were detected using the 561 laser (Yellow/Green) 660/20 filter for BacLight red and gated using the side scatter. CytoPhase+ cells were detected with the 355 (UV) laser and 450/50 filter.

Confirmation of host depletion (HD) strategy. The efficacy of the combined treatment was verified by qPCR in DNA purified from a mixed cell suspension, consisting of 1 x 10^7 *E. coli* cells and 1 x 10^4 4T1 cells. Cells were incubated for 30 min at 37ºC, shaking at 360 rpm in TBS or the optimised **HD buffer** (0.2% Saponin, 0.2% Tween-20, in TBS + $MgCl_2$ (20mM Tris-HCL, 20 mM NaCl, 2mM $MgCl_2$), pH 8) with or without 500 U of Benzonase. The treated cells were then processed for DNA purification using the QIAGEN FFPE kit and the purified DNA analysed by qPCR.

4. Bacterial lysis strategy

Membrane/Cell wall Disruption. FFPE bacterial cells, fixed for 48h, to numbers specified in results, the deparaffinised contents off FFPE slides were treated with 200 µl of PBS +/- 150 µg of Metapolyzyme (Sigma-Aldrich, MAC4L) for 4h. After which cells were either processed for qPCR using the QIAGEN FFPE DNA kit (Qiagen Inc., Valencia, CA, USA). *B. longum* cells FF for 48h were treated with 200 µl of PBS +/- 150 µg of Metapolyzyme (Sigma-Aldrich, MAC4L) for 1, 4 or 24h. After this, contents from each reaction were split in 2. Half the contents were DNA purified for qPCR and the other half was stained with BacLight and SytoBC for flow cytometry analysis. BacLight stains the bacterial cell wall, so BacLight+ bacteria were Wall+. These were detected with 561 laser (Yellow/Green) 660/20 filter. SytoBC stains bacterial DNA, and SytoBC+ cells were DNA+ cells. These were detected with the 488-1 (FITC) 525/50 filter. Cells were gated using the side scatter, and only cells positive for both dyes were considered to be integral. 100,000 events were recorded per replicate analysed. Cells were counted using counting beads as described for in methods section I.a. Counting bacterial cells.

Sample digestion. Optimisation of sample digestion was performed using 8 x 12 µm slides from FFPE blocks loaded with 1x 10^8 *E. coli* cells. The slides were deparaffinised using Zymo deparaffinization solution and resuspended in 200 µl of Lysis Buffer, supplemented with 20 µg of Proteinase K (Qiagen), a top-up of Proteinase K was done every 18 h until reaching the incubation time-point. After tested incubations, DNA was purified using the QIAGEN FFPE DNA kit. 5 µl of purified DNA were loaded to a qPCR reaction for quantitative analysis.

5. Integration of the protocol

Filtering columns. Sterile filtering columns were sought as means to perform the several treatments and washes required for processing the samples. The membrane material for the columns for which results are shown here are either Hydrophilic Polyvinylidene Fluoride (PVDF) (Ultrafree-MC Centrifugal Filters, Merck Millipore) or a Cellulose Acetate (CA) (Corning Costar Spin-X Centrifuge tube filters, Sigma-Aldrich), with pore sizes of 0.1 µm or 0.2 µm. Corning Costar 2 ml microcentrifuge tubes (Sigma-Aldrich, CLS3213) fit CA columns and were used as collection tubes for these columns. Sarstedt 2 ml PP tubes (Sarstedt, 72.689) were found to fit PVDF membranes, and were used as collection tubes for these columns.

Deparaffination. FFPE slides loaded to a filtering column (0.2 µm, CA filter) were deparaffinated by heating the suspension for 2 min at 56°C with 500 µl of Zymo deparaffination solution (Zymo Research, Irvine, CA, USA). The solution was removed by centrifuging the column for 1 min at 1,000 x g.

Testing the removal of enzymes from samples. The ability of filtering columns to allow the removal of proteins > 60 KDa was tested in 0.1 µm PVDF filters. Bovine Serum Albumin (BSA) was selected as a model protein. 200 µl of bacteria suspension with 250 µg/ml BSA in TBS (pH 7.6) were passed through a filtering membrane by centrifuging at 2,000 x g for 1 min, the filtrate (FT) collected and 200 µl of TBS were added to the column, mixed thoroughly and saved as elute fraction (E). This process was repeated twice, until 3 FT were collected. 15 µl of each fraction were loaded to a 4-12% graduated SDS-PAGE gel (Invitrogen) and run for 45 min at 200V. The gel was resolved using EZ Blue (Sigma-Aldrich) and imaged and quantified with a Gel Doc EZ System (Bio-Rad).

Testing adaptation & integration of strategies. Experiments for integrating and adapting the steps of the protocol were performed using 8 x 12 µm slides from FFPE blocks loaded with 1x 10^8 *E. coli* and *S. aureus* cells. Sections were loaded into a 0.2 µm CA filter column (unless specified), deparaffinated as described in section b, and subjected to experimental conditions. Unless specified, the wash buffer is TBS buffer (50 mM Tris, 150 mM NaCl, pH 7.6). All solutions were filtered by a 1 min centrifugation at the specified speed. 100 µl of eluates were collected after each treatment from each replicate and were pooled into one tube that was processed for

DNA purification. Filtrates were collected at the end of the protocol, resuspended in 200 µl. DNA purification was performed using the QIAGEN FFPE DNA kit. Purified DNA was eluted in 40 µl of 10 mM Tris-HCl. A 5 µl aliquot was amplified and resolved using methods described in conventional PCR and run on an agarose gel as described in Methods section II.

Impact of all process and filter pore size on bacteria. The loss of bacterial cells in filter columns was evaluated on 0.1µm PVDF and 0.2 µm CA filters. FFPE slides, rehydrated and deparaffinated, were washed twice in a PBS solution. Aliquots of filtrates and eluates DNA purified and analysed through conventional PCR.

Incorporating tissue rehydration. Deparaffinated slides were rehydrated with 3 washes: 1 x 400 µl wash of 100%, 1 x 400 µl of 80% ethanol and 1 x 400 µl of TBS. Once the solutions were loaded, the mixture was briefly mixed by vortexing and removed by centrifugation at 800 x g for 100% ethanol, 1,200 x g for 80% ethanol and 1,400 x g for TBS. Rehydrated slides, were then treated with HD solution (without DNAse) for 30 min at 37ºC, after which they were washed with PBS. Slides that were not rehydrated were directly resuspended in the HD solution, and all other downstream procedures were the same as rehydrated slides.

Adapting the HD strategy to filtering columns. A time-point incubation with 300 µl of HD solution (without DNAse) for 10, 20 or 30 min, was performed on deparaffinated and rehydrated sections. The reaction was stopped by the addition of 50 mM EDTA and the solution removed by centrifuging at 1,800 x g. The cells washed with 300 µl of TBS, filtered by centrifugation at 1,800 x g for 1 min and the final eluate resuspended in 200 µl of TBS.

Incorporating a tissue dissociation strategy. The effect of tissue dissociation solutions on bacteria was explored with Proteinase K (Qiagen) and Collagenase/Dispase (Sigma-Aldrich). Deparaffinised sections were incubated with 300 µl with 20 µg of Proteinase K for 10 min at 56ºC or 5 µg/µl of Collagenase /Dispase for 1 h at 37ºC. The solution was filtered by centrifuging at 1,600 x g, contents washed with 300 µl of TBS, centrifuged at 1,600 x g, and then incubated with HD solution (without DNAse) for 30 min at 37 ºC. This was removed by centrifuging at 1,800 x g.

6. Validation of the protocol

A. Protocol for bacterial DNA isolation from FFPE samples

1. Tissue sectioning:

Materials:

- Corning Costar 0.2 µm CA filter columns [1 per experimental replicate]
- DISTEL disinfecting wipes (Tristel Solutions Ltd., Cambridge, UK)
- DNA decontamination solution (Sigma-Aldrich)
- Milli-Q water
- Microtome blades [1 per FFPE block]
- Sterile Petri-dishes
- Sterile forceps, individually wrapped [1 per FFPE block]
- Sterile scalpel, individually wrapped [1 per FFPE block]
- Sterile gloves [1 pair per FFPE block]
- Mask

Before starting: the microtome and cold plate are thoroughly cleaned with DISTEL microbial/DNA/RNA disinfectant wipes and a DNA decontaminating solution (Sigma). FFPE blocks are wiped with DISTEL wipes and placed in a cold plate.

Notes:

(i) This protocol was validating using 0.2 µm CA sterile filtering columns.
(ii) To maintain an aseptic technique a mask and sterile gloves, are worn at all times.
(iii) Use a new microtome blade per block sectioned, to avoid cross-contamination.
(iv) Slides a handled with individually wrapped, sterile forceps, only. Use a new pair of forceps per block.
(v) Sterile gloves were changed for handling different blocks.
(vi) Wiping was performed using DISTEL disinfecting wipes.

(vii) If sections from the same block will be taken for staining, cut the sections for staining after cutting the sections for DNA purification have been cut and sterilize probes and brushes to be used by autoclaving and wiping them with a DNAse decontaminating solution.

(viii) If working with difficult to cut tissue that requires rehydrating. Aliquot 20 ml of Milli-Q water in a sterile petri dish (per block), keep it covered until used.

Procedure:

Slides are cut when the equipment is disinfected, the blocks were cooled and all material to be used per-block (sterile forceps, sterile-scalpel, labelled filtering columns, blades, wipes, gloves and mask) placed at hand to avoid touching other surfaces.

(1) A new disinfected blade is placed in the microtome.

(2) An FFPE block is taken and with a new sterile scalpel a small amount of wax from a corner scraped and placed in a filtering-column labelled as wax control.

(3) The FFPE block is placed in the microtome and carefully aligned to the knife's edge.

(4) The microtome trimming thickness is set to 10 µm and 30 –50 µm of the block are trimmed to achieve a full face.

(5) With the block still held in the microtome, any residual wax is wiped off the microtome, and the cutting surface and blade carefully wiped. The blade is repositioned so that a new part of the blade cuts the sections for DNA purification.

(6) The microtome thickness is set to 10 – 12 µm.

(7) Sections are cut slowly, letting them roll and grabbed with forceps before falling from blade/cutting area.

(8) 3 – 5 sections are placed into each replicate sterile 0.2 µm CA membrane filtering column.

(9) After sectioning each block, used blades are removed, residual wax material in microtome disposed and the surface of the microtome disinfected.

(10) Gloves are changed and steps 1-9 were repeated for each block.

1. Dewaxing and tissue rehydration:

Consumables & Solutions (per reaction):

- 4 collection tubes
- 450 µl of Zymo deparaffination solution
- 450 µl of 100% Ethanol
- 450 µl of 80% Ethanol
- 300 µl of TBS (50 mM Tris, 150 mM NaCl, pH 7.6)

Notes:

(i) All present and downstream procedures must be performed in a laminar flow hood.
(ii) All present and downstream material must be UV sterilised before.
(iii) All present and downstream solutions must be prepared and sterilised previously.

Procedure:

(1) Add 450 µl of deparaffination solution into each column.
(2) Incubate at 56°C for 2 min.
(3) Centrifuge for 1 min at 1,000 x g. *[Note: If there is remaining wax – repeat step 1 -3]*
(4) Discard the collection tube and place the filtering column in a new collection tube.
(5) Add 450 µl of 100% ethanol, briefly mix by vortexing.
(6) Centrifuge at 1,000 x g for 1 min, discard the collection tube and place the filtering column in a new collection tube.
(7) Add 450 µl of 80% ethanol, briefly mix by vortexing.
(8) Centrifuge at 1,200 x g for 1 min, discard the collection tube and place the filtering column in a new collection tube.
(9) Add 300 µl of TBS, mix thoroughly by vortexing
(10) Centrifuge at 1,200 x g for 1 min, or until the solution has completely flowed through.

(11) Discard the collection tube and place the filtering column in a new collection tube.

2. Tissue dissociation and host depletion:

Consumables & Solutions (per reaction):

- 5 Collection tubes
- Tissue dissociation solution (TDS): 250 µl of TBS (50 mM Tris, 150 mM NaCl, pH 7.6) supplemented with 20 µg of Proteinase K (Qiagen)
- Host Depletion Buffer (HDB): 300 µl of TBS + $MgCl_2$ buffer (20mM Tris-HCL, 20 mM NaCl, 2mM MgCl2, pH 8)
- Host Depletion Solution (HDS): 300 µl of a solution with 375 U Benzonase, 0.2% Saponin, 0.1% Tween-20, in TBS + $MgCl_2$ (20mM Tris-HCL, 20 mM NaCl, 2mM $MgCl_2$), pH 8)
- Stop Solution: 20 µl of EDTA (500 mM) for a final ~ 33 mM EDTA concentration.
- 400 µl Bacterial lysis buffer = PBS (1X)

Procedure:

(1) Add 250 µl of TDS, mix thoroughly by pipetting. Incubate at 56^{o}C for 5 - 8 min (tissue is noticeably dissociated), shaking at 800 rpm.
(2) Centrifuge at 1,500 x g for 1 min, or until the solution has completely flowed through.
(3) Discard the collection tube and place the filtering column in a new collection tube.
(4) Wash off any remaining Proteinase K with 300 µl of TDB, mix thoroughly by pipetting to achieve a uniform suspension.
(5) Centrifuge at 1,500 x g for 1 min, or until the solution has completely flowed through.
(6) Discard the collection tube and place the filtering column in a new collection tube.
(7) Add 300 µl of HDS, mix thoroughly by pipetting to achieve a uniform suspension and incubate at 37^{o}C for 15 – 20 min, shaking at 650 rpm.

(8) Stop Benzonase activity by adding 20 µl of stop solution. Mix thoroughly by pipetting.
(9) Centrifuge at 1,800 x g for 1 min, or until the solution has completely flowed through.
(10) Discard the collection tube and place the filtering column in a new collection tube.
(11) Wash contents with 400 µl of 1X PBS, mix thoroughly by pipetting to make a uniform suspension.
(12) Centrifuge at 2,000 x g for 1 min, or until the solution has completely flowed through.
(13) Discard the collection tube and place the filtering column in a new collection tube.

4. Bacterial and protein lysis:

Consumables & Solutions (per reaction):

- 1 x 2 ml centrifuge tube (labelled)
- 70 µl of Proteinase K (20 mg/ml)
- 200 µl of Bacterial Lysis solution (BLS) = 1X PBS supplemented with 150 – 200 µg of Metapolyzyme.
- 160 µl of Decrosslinking solution (DCL) = (2 M GuHCL, 1.25% Triton X-100, 1.25% Tween-20, 75 mM EDTA (pH 8.0), 100 mM Tris-HCl (pH 8.0)), for a final solution of (800 mM GuHCL, 0.5% Triton X-100, 0. 5% Tween-20, 30 mM EDTA (pH 8.0), 40 mM Tris-HCl (pH 8.0)).

Procedure:

(1) Resuspend contents in 200 µL of BLS, mix contents thoroughly by pipetting slowly several times until achieving a uniform suspension. Incubate for 4h at 35^{o}C, shaking at 460 rpm.
(2) While incubating, add 30 µl of Proteinase K (20 mg/ml) to the bottom of a 2 ml centrifuge tube and label the tube.
(3) Spin the columns briefly to avoid aerosols
(4) Add to the column 160 µl of DCL, and mix thoroughly by pipetting.

(5) Transfer all the contents of the column to the 2 ml tube with Proteinase K.
(6) Incubate for 18 h at 56ºC, shaking at 700 rpm
(7) After 18h, spin the tube briefly to avoid aerosols, and top up the mix with 20 µl of Proteinase K. Incubate for another 18 h at 56ºC, shaking at 700 rpm.
(8) Repeat Proteinase K top up (20 µl) and incubate for further 6 – 12 hours if the lysate is not completely clear.

5. Decrosslinking and DNA purification:

Notes: The procedures described in for this step includes adapted procedures from the QIAGEN FFPE DNA kit protocol

Consumables & Solutions (per reaction):

- 1 x 1.5 ml centrifuge tube (labelled)
- 1 QIAGEN FFPE DNA purification silica based column
- 5 QIAGEN collection tubes
- 700 µL of column binding buffer made with a 1:1 mixture of 100% ethanol and Buffer AL(Qiagen)
- 600 µl of Buffer AW1 (Qiagen)
- 600 µl of Buffer AW2 (Qiagen)
- 50 µl Elution buffer (EB): Tris-HCl (10 mM)

Procedure:

(1) Set a thermoblock at 80oC, when the block has reached the temperature, transfer the lysed samples into the block and incubate for 1 h.
(2) After incubation, remove tubes form block and let them cool for 5 min.
(3) Spin briefly to avoid aerosols.
(4) Add 700 µl of Column binding buffer to the sample and mix by inverting 3-4 times.
(5) Spin briefly to avoid aerosols

(6) Transfer 650 µl of the tube contents to a Qiagen column, centrifuge at 8,000 x g for 1 min, and change collection tube
(7) Transfer remaining sample content into the column and centrifuge at 8,000 x g for 1 min, and change collection tube.
(8) Wash bound DNA with 600 µl of AW1 wash buffer. Centrifuge at 8,000 x g for 1 min, and change collection tube.
(9) Wash bound DNA with 600 µl of AW2 wash buffer. Centrifuge at 8,000 x g for 1 min, and change collection tube.
(10) Centrifuge at 18,000 x g for 3 min, and place column in a 1.5 ml centrifuge tube
(11) Add 50 µl of Tris-HCl (warm at 50^{o}C) and incubate for 5 min at rt
(12) Centrifuge at 18,000 x g for 1 min, remove column and store eluted DNA at 4^{o}C for next steps.

6. DNA repair:

Notes: The procedures described in for this step includes adapted procedures from AMPURE XP magnetic beads DNA clean-up protocol

Consumables & Solutions (per reaction):

- 2 x 1.5 ml tube or 1 well in a 96-well plate
- Reaction buffer to the following concentrations: 1X NEB CutSmart buffer, supplemented with 100 µM of dNTPs, 50 µM of NAD+ and 2 mM of DTT.
- Enzymes per ng of DNA in the reaction, mix: 0.004 U Formamidopyrimidine DNA glycosylase (FPG), 0.010 U of Endonuclease VIII (Endo VIII), 0.016 U of T4 – Polynucleotide DNA Kinase (T4-PNK), 0.014 U of DNA polymerase I (Pol I) and 0.024 U of *E. coli* DNA ligase (Ligase). All enzymes are acquired from NEB.
- For a 40 µl reaction, with 25 µl of template DNA with a total content of 1,000 ng, add: 4 µl of NEB CutSmart Buffer, 0.10 µl of 40 mM dNTPs (NEB), 0.5 µl of 50 mM NAD+ (NEB), 0.08 µl of 1M DTT (Sigma-Aldrich) and 3.42 µl of nuclease-free water. 0.48 µl of FPG, 0.96 µl of Endo VIII, 1.54 µl of T4-PNK, 1.34 µL of Pol I and 2.3 µl of Ligase.
- Agencourt AMPure XP magnetic beads (Beckman Coulter, IN, USA)

- 80% Ethanol (recently prepared)
- Magnetic rack
- Elution buffer (EB): Tris-HCl (10 mM)
- ~ 1,000 ng of template DNA

Procedure:

(1) Measure DNA concentration using a QUBIT and aliquot 1,000 ng from each sample to the reaction tube. If concentration of samples is between 10 – 30 ng, proceed as follow, for higher or lower DNA concentrations, adjust reaction volumes
(2) Prepare a master mix of the buffer by multiplying the number of samples to be treated by the amount of reagent required.
(3) Dispense the volume of master mix (8.10 µl) into each reaction tube or well
(4) Dispense 25 µl of DNA (~20 ng/µl) into the reactions tube
(5) Dispense the enzyme mix [Keep on ice or cooled block until all reactions are ready]
(6) Transfer to a thermoblock and incubate at 37^{o}C for 30 min.
(7) Stop the reactions by adding 2 volumes of Agencourt AMPure XP magnetic beads, and mix beads and DNA by pipetting at least 10 times
(8) Incubate for 5 minutes , place in magnetic rack and incubate for 2-3 min, until beads are bound to walls and solution is clear
(9) Remove supernatant without removing any beads (remove 110 µl).
(10) With tube/ plate in the rack, wash beads with 200 µl of 80% ethanol, incubate for 1 min
(11) Remove ethanol and again wash beads with 200 µl of 80% ethanol, incubating for 1 min
(12) Remove all ethanol contents and let the beads dry for 3 minutes, until no residual ethanol is observable, but without over-drying beads.
(13) Remove tube/plate from rack
(14) Add 35 µl of EB and mix with beads by pipetting 10 times.
(15) Incubate beads with EB for 5 min.

(16) Place tube/plate in magnetic rack and incubate for 2-3 min, until beads are bound to walls and solution becomes clear.

(17) Transfer eluted DNA into a new tube/plate and store at -20°C if not being used immediately.

B. Bacterial DNA isolation from flash frozen samples

Breast tumour core-biopsies were aseptically resected using an Achieve 14G Breast Biopsy System (Iskus Health, UT, USA). The specimens were transported in sterile PBS to the lab, where they were flash-frozen and kept at -80°C until further processing. DNA from the specimens was purified following the protocol and reagents provided in the Ultra Deep Microbiome Prep (Molzym, GmbH & Co. KG., Bremen, Germany) and eluted in 100 µl of Tris-HCl.

7. Bioinformatics

The quality of the paired-end sequence data was initially visualised using FastQC v0.11.6, and then filtered and trimmed using Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs) and their distribution among the samples. It is recommended that ASVs (formerly called 'Ribosomal Sequence Variants') be used in place of 'operational taxonomic units' (OTU), in part because ASVs give better resolution than OTUs, which are clustered based on similarity. ASVs were then exported back into Linux and a second stage of chimera removal was carried out using USEARCH v9 in conjunction with the ChimeraSlayer Gold database v6 as a relatively high number of cycles was used during PCR amplification.

The following statistical analyses were carried out in R: Shannon alpha diversity and Chao1 species richness metrics, and Bray-Curtis distances, for analysis of beta diversity, were calculated using the PhyloSeq package v1.24, and the Vegan package v2.52. Beta diversity calculations produce distance matrices with as many columns and rows as there are samples; thus, beta diversity is often represented using some form of dimensionality reduction, in this case, using principal co-ordinates analysis (PCoA) with the Ape package v5.1. Hierarchical clustering, an unsupervised method that can reveal key taxa that distinguish their respective environments, was performed with the heatplot function in the made4 package v1.54. Differential abundance analysis was carried out using Deseq2 v1.2.0, which identifies differentially abundant features between two groups within the data. Tests of means were performed using the Mann-Whitney *U* test unless otherwise stated, and correlations were calculated using Spearman's rank correlation coefficient. Where applicable, false positive rates were controlled below 5% using the FDR procedure. Random forest classification trees were run using the RandomForest (v4.6.15) and pROC (v1.15.3) packages in R.

Bioinformatics contamination control. Despite not identifying the contaminant taxa themselves, the source tracker utility is invaluable in estimating the proportion of a sample ("Sink") that may have originated in a negative control ("Source") Decontam can remove taxa, based on presence or absence in negative controls, or inverse correlations with input DNA. Requires a threshold to be set, which can be dictated by SourceTracker. The effectiveness of this can then be confirmed by SourceTracker.

RESULTS

1. Strategy for host DNA depletion

Host membrane permeabilisation. The state of membrane permeabilisation in both, Gram negative bacteria and mammalian cells was evaluated and a permeabilisation agent able to induce pores in mammalian, but not in bacterial cells was investigated. For this, a cell internalisation marker with similar dimensions to that of DNAse was exploited to report on the degree of permeabilisation of mammalian or bacterial cells. DNAse I is a compact monomer with a MW of ~ 30 KDa and dimensions of 4.6 x 4 x 3.5 nm, and most DNAses commercially available have similar dimensions [51]. SAv is a globular tetramer, with a MW of ~ 52 KDa dimension of 5 nm [52]. SAv strongly binds biotin ($K_d \sim 10^{-15}$ M) [53], an intrinsic and essential co-factor for many enzymes in all domains of life, including prokaryotes and eukaryotes [54]. The highly specific, rapid and resistant nature of this interaction has been exploited for many purposes [55]. This includes the detection of naturally biotinylated intracellular molecules, such as proteins [56-59]. With this information, the marker for internalisation used was SAv conjugated with Cy5 (60 KDa) to target any naturally available biotinylated proteins that would be crosslinked due to formalin fixation and therefore would not be lost during cell permeabilisation.

The internalisation of proteins was by flow cytometry in 4T1 cells (mouse breast carcinoma) and *E. coli* as a model organism for G- bacteria. Cell suspensions with 1.25 x 10^6 4T1 cells and 1.25 x 10^7 were exposed to either Triton-X, Tween-20, Saponin, Digitonin or none. Treated cells were then labelled with Streptavidin-Cy5 (SAv-Cy5), as a fluorescent marker for protein internalisation. As seen in Figure 2, impermeabilised 4T1 cells show much less fluorescence than those exposed to a detergent (Top) and only *E. coli* cells treated with Triton-X were permeabilised, as evident from the 363 X ($p < 0.001$) increase in fluorescence. This indicates that pores induced by fixation are not large enough to internalise SAv-CY5, and thus require permeabilisation for allowing its introduction to the cell. From the detergents tested, Saponin showed the highest capacity of permeabilising 4T1 cells, as measured with

the 186 X ($p < 0.001$) increase in fluorescence, without any effects observed in *E. coli* ($p > 0.05$). A further exploration in the concentration of Saponin that can be used without significantly permeabilising *E. coli* cells, allowed for an adjustment in Saponin concentrations used, which was increased from 0.1% to 0.20% (sFigure 1a).

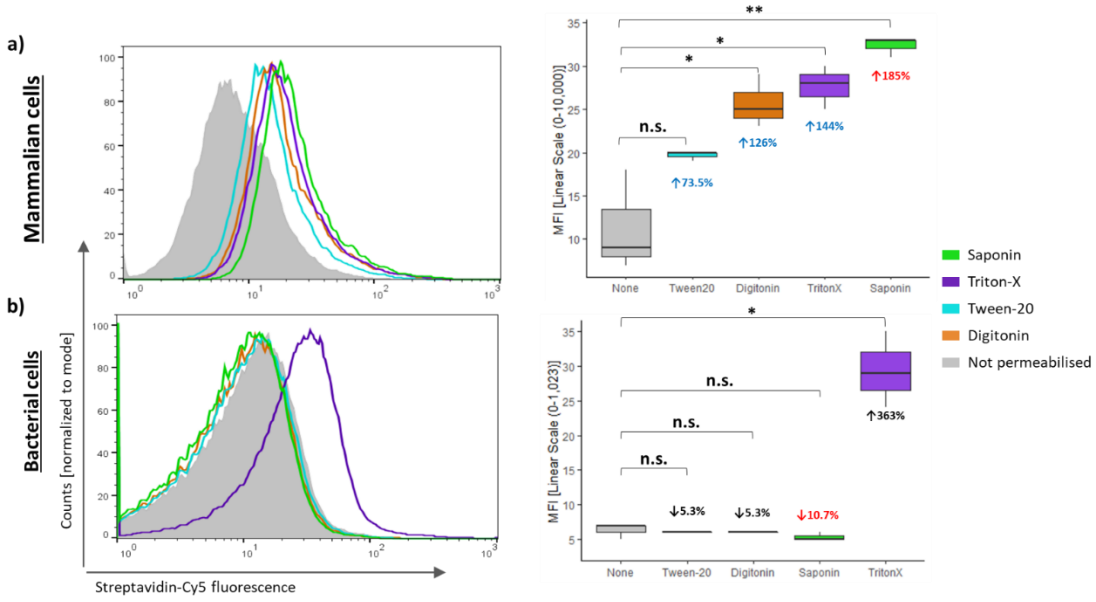


Figure 2. Membrane permeabilisation.

Cell permeabilisation is measured by the internalisation of Cy5 labelled streptavidin (SAv-Cy5 was measured in ***(a)*** *4T1 cells and* ***(b)*** *E. coli.* ***(Left)*** *Histograms showing the maximum fluorescence intensity (Cy5+) after treatment with permeabilisation agents (grey, impermeabilised cells, $n = \bar{x}\ 6$).* ***(Right)*** *Box plot showing median fluorescence intensity of treated and untreated cells (n = 6 for each box). Saponin-treated (green) 4T1 cells show 185 X ($p < 0.001$) increase in SAv-Cy5 intake, while no effect is seen E. coli ($p > 0.05$). n.s. = $p < 0.05$*

<u>*Host DNA depletion.*</u> Among 6 DNAses screened for activity under the conditions set by this sample type and permeabilisation agent, the highest levels of activity was observed for Benzonase (sFigure 1b). Host depletion was tested here by measuring the fluorescence emitted by a DNA intercalating dye (CytoPhase Violet), after treatment with a permeabilisation agent and Benzonase. To ensure that the pore sizes created by permeabilisation agents allowed for the internalisation of Benzonase in 4T1 cells, the 4 permeabilisation agents were tested. For Bacteria, only Saponin was tested and Triton-X was used as a permeabilisation+ control. Here, a reduction in CytoPhase fluorescence of the cell population tested is indicative of a reduction in the DNA available for binding CytoPhase, suggesting a higher activity of DNAse. Results

shown in Figure 3 reflect those in Figure 2. Again, Saponin was the most effective permeabilisation agent for 4T1 cells *(top)*, where a 30.8% ($p < 0.001$) reduction in fluorescence is observed. The opposite was observed in *E. coli (bottom)*, where treatment with Saponin did not show any significant change in fluorescence (4.5% decrease, $p > 0.05$)) from the *Impermeabilised + DNAse* or *Impermeabilised – DNAse controls*. The analysis of bacterial cells was validated with Triton-X where a 43.7% ($p < 0.001$) reduction in fluorescence was observed. Therefore, Saponin + Benzonase were chosen as the reagents for host depletion strategy. An enzyme dose optimisation was performed, where units ranged from 75 U to 375 U (data not shown). Therefore, from this point onward the active components of the Host Depletion (HD) solution were 0.20% Saponin and 375 U of Benzonase.

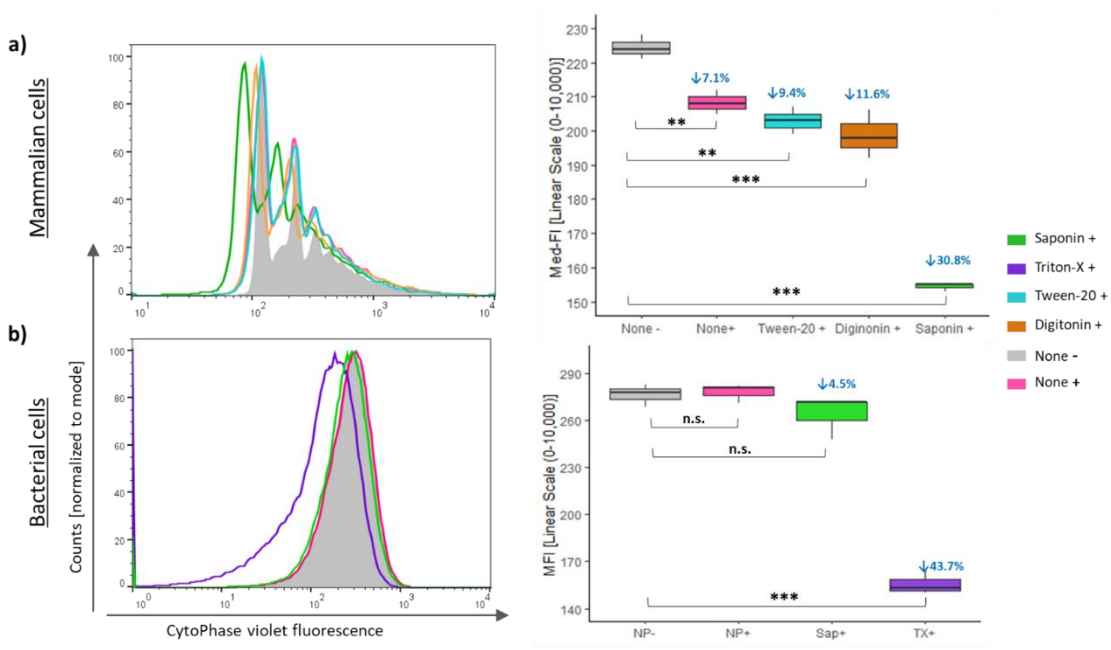


Figure 3. DNA depletion.

DNA depletion is measured here by a reduction in fluorescence of CytoPhase (intercalating DNA dye), after treatment with a permeabilisation agent and Benzonase measured in ***(a)*** *4T1 cells and* ***(b)*** *E. coli.* ***(Left)*** *Histograms showing the maximum fluorescence intensity for CytoPhase+ cells. In grey = impermeabilised, DNAse negative controls, pink = impermeabilised, DNAse+ controls, for each line, $n = \bar{x}$ 6.* ***(Right)*** *Box plot showing median fluorescence intensity of treated and untreated cells (for each box, n = 6). Saponin treated (green) 4T1 cells show 30.8% ($p < 0.001$) decrease in CytoPhase fluorescence, while no effect is seen in E. coli exposed to Saponin and DNAse ($p > 0.05$). Triton-X (Purple) as a Permeabilisation + DNAse treatment control for bacteria, shows a 43.7% ($p < 0.001$) decrease in E. coli cells.*

Quantifying Host DNA depletion. The effect of the HD strategy in terms of DNA depletion was quantified by qPCR, in a mixed FF cell population with 1 x 10^7 *E. coli* and 1 x 10^6 4T1 cells. Here, FF cells exposed to the HD strategy were harvested and DNA purified and eluted DNA analysed by qPCR. As it can be seen in Figure 4, results are concordant with the previous experiments, where the quantity of genomes retrieved is a log-fold ($p < 0.01$) reduced by treatment with the HD strategy. On the other hand, that allows for a higher representation of bacterial DNA, which after treatment with the HD strategy, exhibits a 3X ($p < 0.01$) increase in the number of genomes recovered. Finally, the impact of incubation time was also assessed and confirmed optimal at 20 min (sFigure 2c).

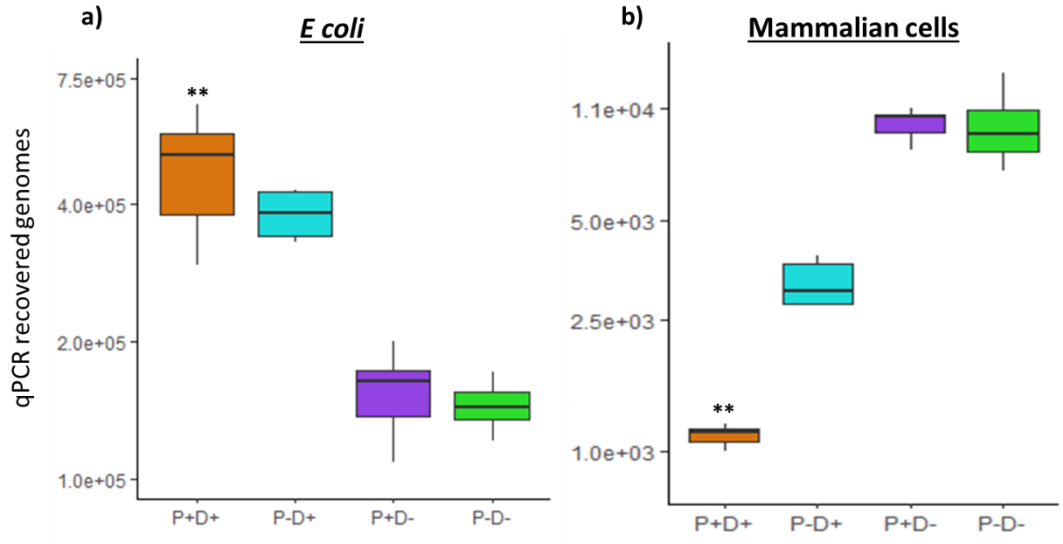


Figure 4. Quantifying Host Depletion.

DNA depletion is measured by a reduction of genomes recovered by qPCR. A mixed cells suspension of ***(a)*** *1 x 10^5 4T1 cells and* ***(b)*** *1 x 10^6 E. coli was treated or not with Saponin and Benzonase, DNA purified and quantified by absolute quantitation. Here,* ***P+D+*** *= Saponin + DNAse + (orange),* ***P-D+*** *= Saponin – DNAse + (cyan),* ***P+D-*** *= Saponin + DNAse – (purple) and* ***P-D-*** *= Saponin – DNAse – (green). For each box, n = 6. As seen here* ***P+D+*** *treated cells show a log fold reduction in the recovery of DNA for 4T1 cells* ***(b)****, and the effect of reduction of host DNA, enriches bacterial DNA in the template, as seen in* ***(a)*** *A 3X increase of E. coli DNA was recovered.*

2. Bacterial lysis strategy

Bacterial lysis in FFPE blocks was examined with a combination of treatments, known to be 'safe' in maintaining the integrity of nucleic acids. These include 1) Membrane/wall disruption with lytic enzymes, and 2) Sample Digestion with chaotrophic agents and Proteinase K.

Membrane/Cell wall disruption. Given the intrinsic DNA damage present in FFPE samples, membrane/cell wall disruption was attempted through a mix of lytic enzymes known to disrupt the walls of gram positive (G+) bacteria and capsids of gram negative bacteria. This was first evaluated by qPCR in FFPE blocks loaded with 1x 10^8 of each: *E. coli*, *S. aureus* and *B. longum* cells, formalin-fixed for 48 h. Deparaffinised contents were incubated in PBS +/- Metapolyzyme for 4 h before DNA purification. qPCR reactions were loaded with the equivalent of 1 x 10^7 bacterial cells of each strain and set to amplify a 200 bp fragment. A marked increase in DNA recovery is evident for all three bacterial strains, indicating that Metapolyzyme lyses both G+ and G- FFPE bacteria (Figure 5a). The effect of Metapolyzyme was more pronounced in *S. aureus*, where treatment with Metapolyzyme increased its recovery by 2-log fold ($p < 0.001$), followed by a 1-log-fold ($p < 0.001$) increase shown for *B. longum* and a 0.5-log-fold ($p < 0.001$) increase observed for *E. coli*. Given the log-fold decrease in the recovery of PCR readable 200 bp long DNA fragments, and that G- cells are more liable to lyse by treatments (centrifugation, exposure to solvents), it can be implied that by incubating these samples with Metapolyzyme more than 70% of the available PCR readable bacterial DNA was recovered. These results were confirmed by a time-point incubation of *B. longum* cells FF for 48 h. 1x 10^9 cells were treated with Metapolyzyme for 1, 4, or 24 h. Half the contents were analysed by flow cytometry (Figure 5bi) and half were analysed by qPCR (Figure 5bii). Results from both analyses are in agreement. The optimal incubation time to recover 'difficult to lyse' *Bifidobacterium* is 4 h at 35^{o}C, shaking at 550 rpm. As can be seen in Figure 5bi, at this point, only 0.25% ($p < 0.01$) of the cells loaded in the reaction were DNA+/Wall+. These results were confirmed by qPCR (5bii), where the highest DNA recovery (71%, $p < 0.001$) of a 121 bp fragment (single genome copy) was observed. On the other hand, 83% of the *Bifidobacterium* cells not treated with Metapolyzyme

after 24 h incubation were still integral (Wall+/DNA+) and from this bacterial population, DNA accessible for purification from only 4.4% of cells. Lastly, upon O/N incubations (24 h) with Metapolyzyme, cells were lysed at the same level as observed for 4 h - however, the DNA contents were reduced by half. This could be an effect of nucleases or DNA denaturing agents released from the crosslinked matrix.

Sample digestion. After membrane rupturing with Metapolyzyme, the complete digestion of the peptidoglycans and other cellular proteins crosslinked to DNA need to be digested. Sample digestion is essential for DNA purification and is included in all DNA purification protocols. However, the choice of denaturing agent is guided by the cell type and macromolecule content of cells. The buffer choice here (chaotrophic), was informed by previous experiments indicating that this buffer has a higher capability of decrosslinking DNA at lower temperatures (*see Ch. 3 - FFPE DNA damage & repair*). Interestingly, Proteinase K is known to have a higher activity in a buffer with similar composition. Sample digestion in FFPE tissue is usually performed with Proteinase K. It has been suggested that longer digestion incubations lead to increase yield of amplifiable DNA [60]. This was confirmed here, where a time point incubation with Proteinase K was performed to inform the incubation length that will lead to the highest yield of amplifiable DNA in FFPE bacteria. Deparaffinised contents of FFPE blocks loaded with *E. coli* were digestested for 1 h, 24 h or 48 h. As can be seen in Figure 5C, longer digestions lead to a higher recovery of long and short DNA fragments, with 1.6X and 2X increase in the recovery of short and long fragments, respectively, after increasing digestion from 1 h to 24 h ($p < 0.001$). A mild 6 % (Short) and 0.5 % (long) increase is observed after increasing the length of incubation from 24 h to 48 h, indicating that after 24 h of incubation with Proteinase K, almost all the protein content of the cell is digested (Figure 5c).

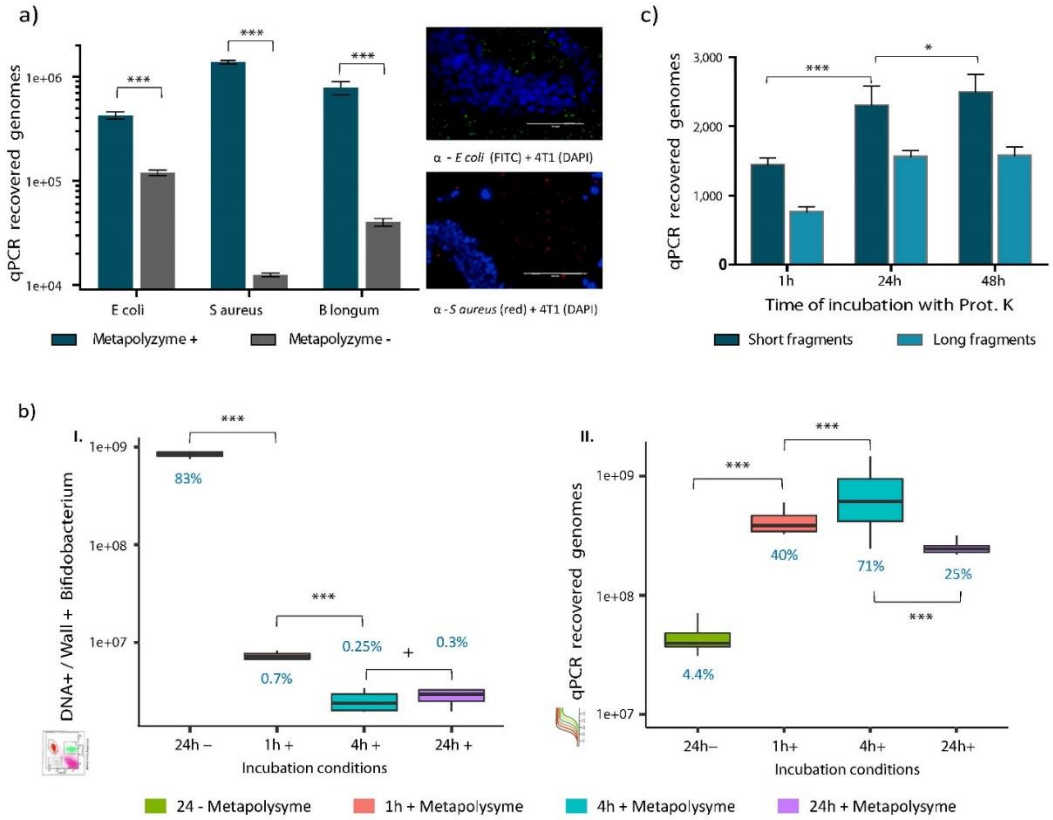


Figure 5. Bacterial Membrane/Cell wall disruption

(a) *<u>Metapolyzyme lyses FFPE bacteria.</u> Bar Plot showing qPCR DNA recovery of a 200 bp, single copy DNA fragment after lysis (blue) /no lysis (grey) with Metapolyzyme. Recovery (black) was estimated from a 1 x 10^7 genome load. For each bar, n = 6. For the 3 strains tested, treatment with Metapolyzyme markedly increased the recovery of DNA (p <0.01). This was more noticeable for S. aureus, which without treatment with Metapolyzyme, only 1 % of the bacterial DNA was detected.*

(b) *<u>Optimising lysis with Metapolyzyme in 'difficult to lyse bacteria'.</u> 1 x 10^9 Bifidobacterium cells incubated with Metapolyzyme at 1 h, 4 h, and 24 h and without Metapolyzyme for 24 h. (i) Box plot showing average cell counts (n = 6) of integral cells (Wall+, as measured by fluorescence of BacLight and DNA+, as measured by fluorescence by SytoBC) after incubation with/without Metapolyzyme. (ii) qPCR recovered genomes recovered from an equal fraction of populations treated (for each box, n = 6). For cells not treated with Metapolyzyme after 24 h (green), the cell count was almost the same as the initial population. Higher lysis levels / DNA recovery was achieved after 4 h incubation (cyan), with only 0.25 % (p<0.001) maintaining integrity and genomes corresponding to 71 % qPCR recovered.*

(c) *<u>Optimising protein lysis.</u> Deparaffinised contents from 8 x 12 µm slides from FFPE blocks loaded with 1 x 10^9 E. coli cells were digested for 1 h, 24 h or 48 h. Genomes recovered by qPCR measuring amplification of a 200 (blue) or 500 (cyan) bp fragment is shown. Increasing digestion time from 1 h to 24 h increases the yield of amplifiable fragments in both short (1.6X) and long (2X) (p < 0.001).*

3. Integration of a Protocol for microbiome analysis of FFPE tissues

As a means to integrate this protocol without the significant loss of bacterial cells / DNA, the use of filtering columns with filter pores sizes allowing a rapid flow (less centrifugation), while retaining bacterial contents, was examined. Several commercially available columns were examined to identify the columns that perform best in terms of parameters required for this protocol (high flow-rate, resistance to solvents, low retention, no bacterial loss, and availability of collection tubes).

To evaluate the choice of columns, first the flow-capacity of a lower flow-rate column (PVDF 0.1 µm) was tested in its ability to filter BSA (64 KDa) from a bacterial suspension. As can be seen in the protein gel in sFigure 2a, after a single centrifugation at 2,000 x g for 1 min, 99.6% of the 250 µg /ml BSA was in the eluate and the remaining 0.4% was completely filtered after a second wash. No protein was detected in a third eluate. On the other hand, filtrates were also subjected to flow cytometry and no bacteria were found in the solution (data not shown), thus confirming that filtering units are a good strategy to remove enzymes from a sample without the need of high speed centrifugation.

After this, the effect of filter pore size of 0.2 µm Cellulose Acetate (CA) filters on bacterial loss was evaluated by PCR of eluates and filtrates of FFPE slides washed twice with PBS. As seen in the gel in sFigure 2b, amplification products for *E. coli* and *S. aureus* were only detected in the filtrate of both 0.1 µm PVDF membranes and 0.2 µm CA filters. The use of 0.2 µm pore CA filter membrane (Corning Costar Spin-X) columns allowed for the entire workflow to be carried out at centrifugation speeds lower than 2,500 x g, for 1 min (see Methods – Protocol).

Once the filters to be used were confirmed, the effect of rehydration and wash of traces of organic residues was assessed. As seen in the gels in sFigure 2c, only in the filtrates of tissues that were rehydrated and washed was there a consistent band for both *E. coli* and *S. aureus*. Conversely, on dehydrated tissues, not all samples amplified, and off-target effects were observed for Saponin. These data mirrored the incubation conditions for host depletion (sFigure 2d), for which an off-target effect that did not affect the output of the assay, was observed for *E. coli* across all incubation time-

points, with lower effects observed in 10 min incubation. Given that, in the models used, bacteria are more readily exposed to the treatment and that during the process of making the model some bacterial membranes can be damaged, it was decided to maintain the host depletion incubation at 20 min.

Finally, two tissue dissociation strategies were evaluated. As can be seen in sFigure 2E, both strategies have an effect in G- bacteria, with higher loss of *E. coli* DNA observed for Collagenase/Dispase, which clearly lysed a portion of the bacteria. For Proteinase K, a band was found after permeabilisation, indicating that Proteinase K debilitates the OM, exposing the phospholipid bilayer, which can be then accessed by Saponin. Since -dissociation of a tightly packed and hardened FFPE tissue is necessary for enzyme access, it was opted for the Proteinase K alternative.

4. Validation of the protocol by 16S sequencing

Assessment of contamination introduced. The level of processing required when creating a sequencing library from FFPE samples, coupled with the anticipated low biomass of the samples, makes them highly susceptible to contamination. Figure 6 shows a representative sample from each sample group, and each library preparation method. The “In House methods” are consistently more susceptible to contamination than the gold standard Qiagen method, and a controllable level of contamination is present in all sample types with the exception of FFPE breast samples, which are overwhelmingly contaminated. The output of the SourceTracker algorithm also indicates which negative controls were implicated in the contamination, and Figure 6a, shows the composition of these samples at the family level.

The use of Protoblocks with known bacterial composition allowed for accurate quantification of the number of sequencing reads lost due to contamination between the three different treatment groups. As seen in Figure 7, in both the Qiagen FFPE protocol (Q) and the *In house with host depletion* (IHP) protocols, the proportion of reads removed as part of the contamination control workflow was less than 5%. With the *in house without host depletion* (IHN) protocol, almost a quarter of all reads obtained for these samples had to be removed.

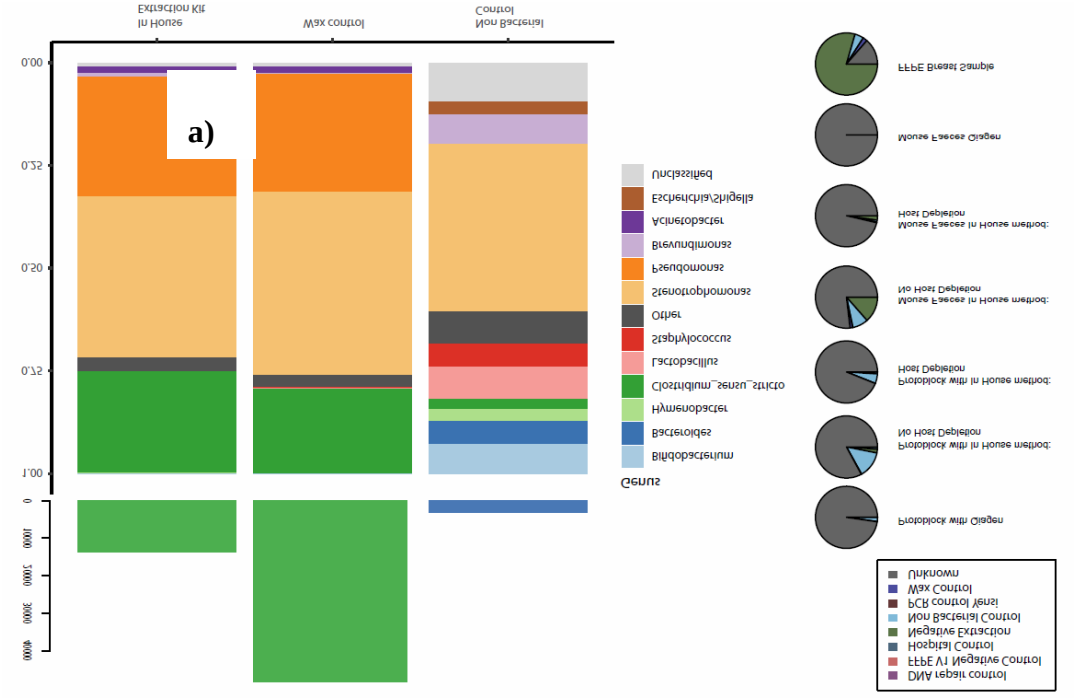


Figure 6. Inherent contamination present in method.

(a) *Shows the sample composition of the three negative controls implicated by the SourceTracker algorithm.*

(b) *Shows the output of the SourceTracker algorithm, with one representative pie chart per sample type indicating the degree of contamination present.*

Data indicate that although contamination is present in most samples only FFPE breast samples are overwhelmingly affected by environmental contaminants. In addition, only 3 of the 8 negative controls are implicated.

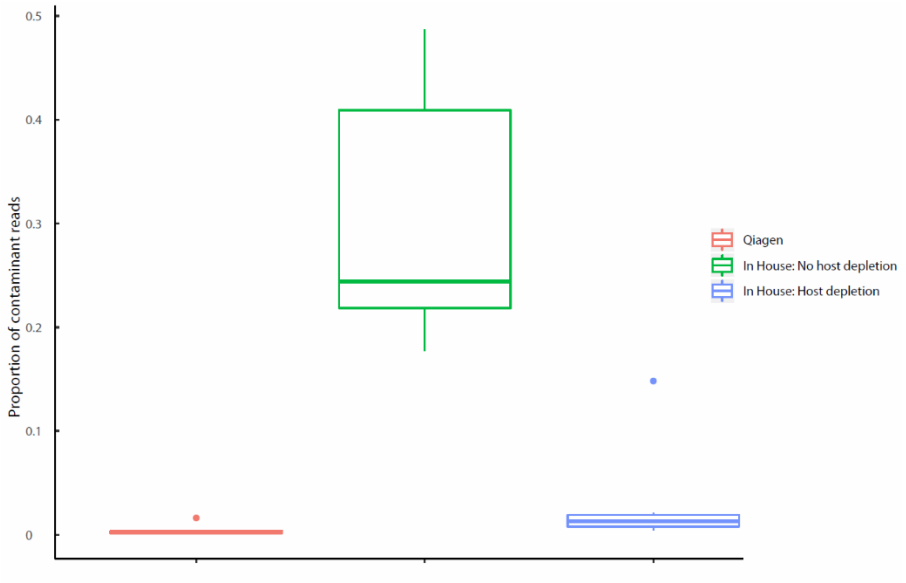


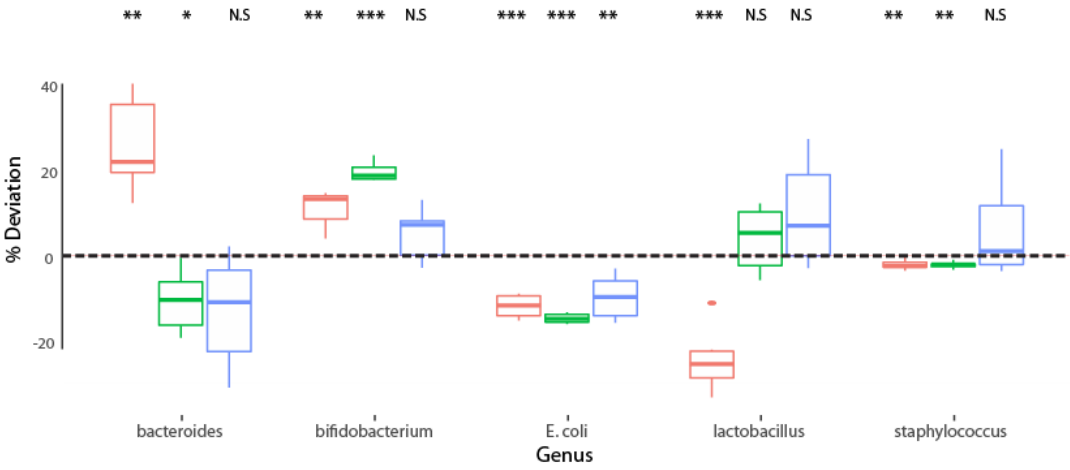
Figure 7. Proportion of reads lost due to environmental contamination introduced during processing.

The data indicates that while only a marginal percentage of reads are consumed by environmental contaminant DNA in the Qiagen and IHP samples, just under 30% of reads on average are lost in IHN samples.

Assessment of method in protoblocks Samples labelled as IHP went through the DNA protocol (bacterial lysis, sample digestion, DNA purification and repair) plus a host depletion step, and IHN samples, did not include a host depletion step. The precise quantities of bacteria added to the FFPE mock communities can be seen in Supplementary Table 1. This information allowed a robust analysis of methodological bias in terms of under or over-representation of different bacteria.

As shown in Figure 8, the Q protocol, which is not optimised for bacterial DNA, showed statistically significant under or overrepresentations in all 5 genera present in the Protoblock, particularly in the case of *Bacteroides* and *Lactobacillus*, which were over and underrepresented by more than 20% respectively. In the IHN method, no significant bias was observed with lactobacillus, the deviation in *Bacteroides* was marginally significant, while all other genera were significantly under or overrepresented. The IHP method was the least susceptible to bias, with only the proportion of *E. coli* presenting as significantly different from what was theoretically present in the Protoblock. These findings are supported by qPCR recovery analysis of the 5 bacterial species added to the block, seen in Supplementary Figure 3.

a)



b)

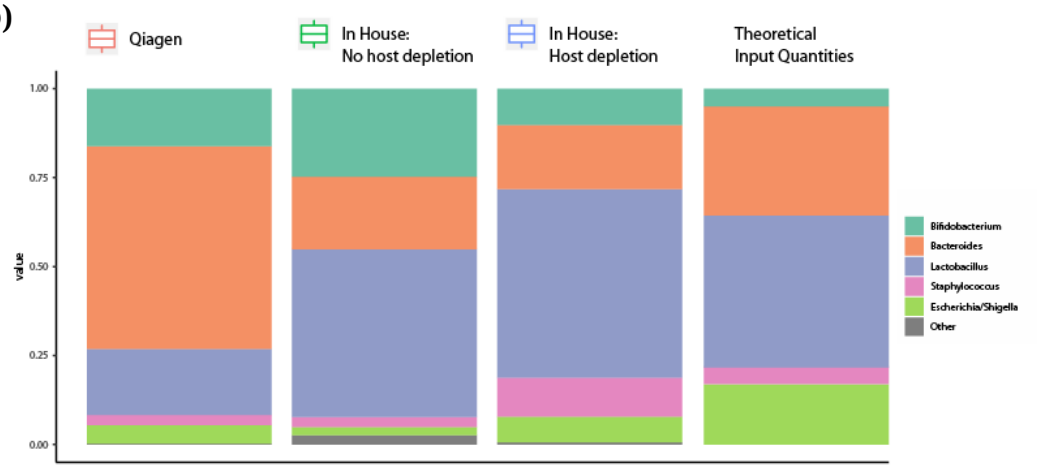


Figure 8. Assessment of bias in terms of bacterial community composition between methods.

(a) *Shows the percentage deviation of bacterial composition per genera, per extraction method, from the original quantities input into the protoblock.*

(b) *Shows sample composition of all samples merged by extraction kit, with the right most column representing the ideal proportions as dictated by the input quantities.*

Visually IHP has the least degree of bias over the five bacterial genera. This is confirmed statistically in ***(a)****.*

Assessment of method in murine models. The comparisons facilitated by the protoblocks were complemented by mouse faecal samples, which were formalin fixed

and paraffin embedded as described in methods (murine models) and their protocol included bacterial lysis, sample digestion, DNA purification and DNA repair, with host depletion + tissue dissociation (DT-P) or without any of these 2 treatments (DT-N). The community structure in these samples was considerably more complex than in the Protoblock.

Beta diversity analysis using Bray-Curtis dissimilarity shows no significant difference between the IHN and Qiagen methods. This can be seen visually as the samples cluster together, and is confirmed by PERMANOVA analysis, (p = 0.231). Both Qiagen and DT-N are significantly dissimilar to DT-P as per PERMANOVA, (p = <0.001) (Figure 9).

The driving factors behind the distinct clustering were assessed by searching for correlations between the dominant bacterial families seen in the samples, and either of the two principal coordinate axes. The correlations were carried out using Spearman's method, and multiple testing was controlled for using the FDR method. This was expanded upon in Figure 10, with a direct comparison of sample composition between FFPE vs Flash-frozen samples in the three treatment groups. Figure 10a compares Q and DT-P paired samples. In this instance, the Gram positive *Coriobacteriaceae* and *Lactobacillaceae* were significantly elevated in terms of mean proportion in the DT-P samples, while the Gram negative *Porphyromonadaceae*, *Rickenellaceae*, *Prevotellaceae* and *Bacteroidacaeae* were elevated in samples treated with Q. Figure 10b compares the paired samples prepared using the Q and DT-N methods respectively. In this instance, the there was no significant difference in the Gram positive families, while the two previously indicated Gram negative families *Coriobacteriaceae* and *Lactobacillaceae* were elevated in the DT-N group. Also elevated were the *Pseudomonadaceae* and *Promicromonosporaceae* families, which are likely to be residual environmental contaminants missed by the retrospective bioinformatic contamination removal. Figure 10c compares the in house method with and without *host depletion + tissue dissociation*, where the difference was in the Gram negative families, which were elevated in the DT-N samples.

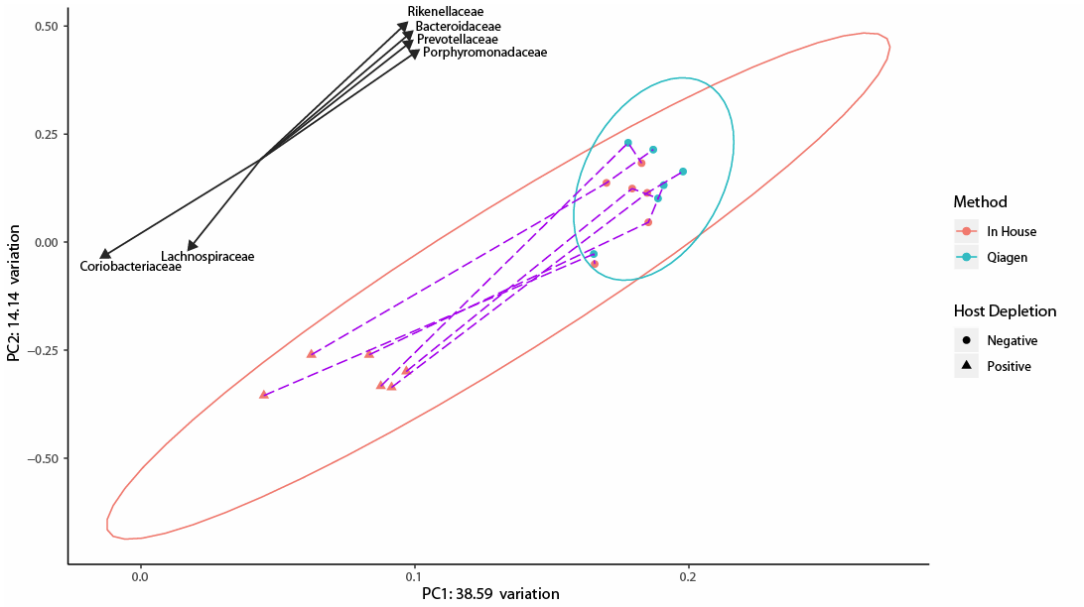


Figure 9. Principal Coordinate analysis of matched murine samples.

Points coloured by extraction method, and shaped by host depletion status. PcOA plot supported by correlations of major bacterial families present in dataset with PC1 and PC2 values used to generate plot. Only significantly correlating families show, with significance tested for using Spearman's method. False discovery rate controlled for using FDR method. Data indicates that host depletion strategy has an effect on Gram negative bacteria.

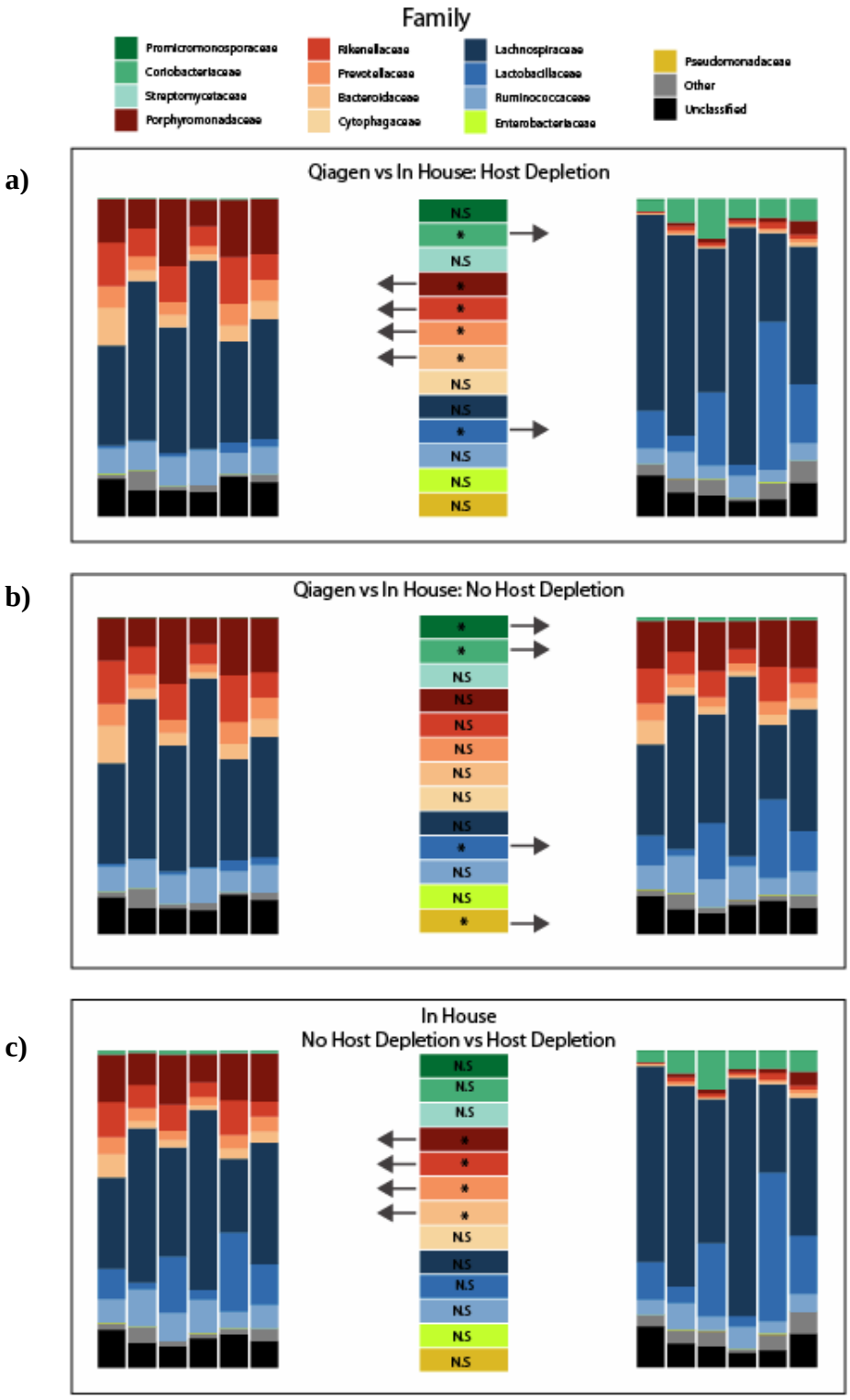


Figure 10. Mouse faecal sample composition comparison between methods.

Mean abundance of major families between groups tested using Wilcox signed rank test, with false discovery rate controlled for using the FDR method. The arrow indicates the direction if increase in cases of significant difference. ***(a)*** *Compares Qiagen with IHP.* ***(b)*** *Compares Qiagen with IHN.* ***(c)*** *Compares IHN with IHP.*

Assessment of protocol in Patient samples. The final assessment of the method was the analysis of FFPE malignant breast tissue samples. The accuracy was verified by comparing the FFPE samples with their matched freshly frozen samples. As was suggested by the representative pie chart of the FFPE breast samples in Figure 6b, the quantity of environmental contamination was overwhelming, this was unsurprising given the low level of microbial biomass present in the samples. Even after contamination removal, leaving all other sample types with little to no contamination, the FFPE breast samples in Figure 11b are dominated by the *Pseudomonadaceae* and *Xanthomonadaceae* families seen in the negative control samples and bear little resemblance to their fresh frozen counterparts in Figure 11C. However, when Figure 11b is recreated in 11d, with all *Pseudomonadaceae* and *Xanthomonadaceae* associated sequences manually removed, there is resemblance between the two groups that begins to justify what the Venn diagram in Figure 11a indicates in terms of shared bacterial families. In total 24.6% of the total bacterial abundance in Figure 11d is accounted for by bacterial families also found in the fresh samples shown in Figure 11b.

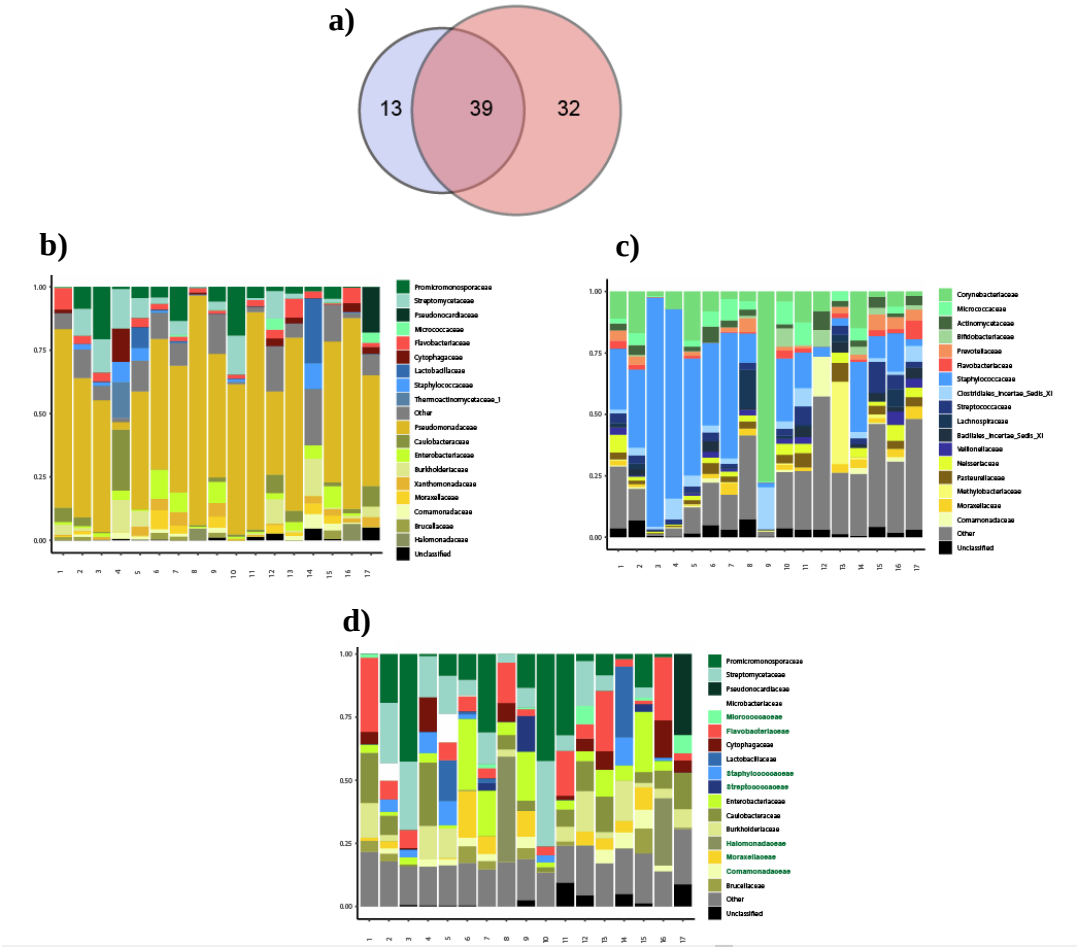


***Figure 11. Sample composition comparison between matched patient samples.* 4**

(a) *Venn diagram visualising the observed families in* ***(b)*** *FFPE breast tissue and* ***(c)*** *Matched fresh frozen breast samples processed using Molzym Ultra-Deep Microbiome kit.* ***(d)*** *FFPE samples with the two obvious contaminant families, Xanthomonadaceae and Pseudomonadaceae manually removed.*

DISCUSSION

Given the potentials that FFPE material could bring to the field of the human microbiome, a method that allows access to this material is essential. Currently, there are no methods available to process this sample type for microbiome studies. This research presents foundational strategies to treat these samples in order to guarantee a truthful representation of the bacterial communities inhabiting tissues.

Bias generated by host DNA was confirmed for FFPE samples by qPCR (sFigure 3A), it was therefore relevant to this project to devise a host DNA depletion assay for FFPE samples. It is well established that fixatives permeabilise eukaryotic cell membranes, allowing the passage of small molecules. Although the exact size of pores induced by fixation has not been investigated, several protocols include permeabilisation steps to enhance the internalisation of larger molecules, such as antibodies. In this study, we explored whether the size of pores induced by fixation allows the entrance of proteins the size of DNAse into cells. Data shown in figure 2, confirms that formalin fixation does not induce pores that are large enough to introduce molecules of 4-5 nm of diameter, while all permeabilisation agents tested here were proven to induce pores >5 nm of diameter in mammalian cells, thus, allowing the internalisation of proteins of the size of DNAse I and Streptavidin. Among the permeabilisation agents tested, the best mammalian cell selective permeabilising agent was Saponin, which exhibited high Cy5+ internalisation in mammalian cells, without having any effect in bacterial cells. Of note, in experiments performed at early stages of this project (sFigu1c), the use of chaotrophes was found detrimental for this sample type.

After confirming permeabilisation, a screen of DNAses, informed the choice for Benzonase as the nuclease to use in this host depletion strategy. While most of the DNAses tested here, are monomers with an average MW of 30 KDa, Benzonase, is a dimeric nuclease, each monomer with a MW of 30 KDa, similar to the size of SAv-Cy5. [61]. Its activity was confirmed by DNA depletion experiments. The results from these experiments were in agreement with results from cell permeabilisation. A marked reduction in DNA quantity, as analysed by flow cytometry and qPCR, was observed after treatment with the complete HD strategy (Saponin + Benzonase) was observed for host (4T1) cells only (Figure 3 & 4). These results are supported by

experiments performed on protoblocks (sFigure 3B) and recently-published (during the course of this project) evidence on non-fixed samples, showing that Saponin outperforms other permeabilisation agents trialled for host-depletion strategies [62, 63]. Similarly, 4other authors have found on non-fixed samples that Benzonase is the most effective DNAse for host depletion strategies [16, 17]. It must be noted that during tissue dehydration, a fraction of cell membrane lipids is lost (up to 40%) [64], it is therefore expected to observe a reduction in the permeabilisation efficacy once applied to FFPE tissues.

Results presented here indicate that FFPE processing does not severely debilitate bacterial cell walls, meaning a bacterial lysis mechanism must be included in sample processing for microbiome analysis. Without this, the results will be biased towards easier to lyse G- bacteria. Bacterial FFPE DNA is inherently damaged and harsh lysis methods would be detrimental, reducing already low yields and further fragmenting an already severely fragmented DNA. Due to time constrains and the complexity of integrating bead beating or other mechanical lysis strategies to the workflow, these were not tested here to set a reference. However, numerous reports have confirmed this effect in different NF sample types and in archaic samples. The later having a DNA damage profile resembling that of FFPE DNA [40, 65-68]. Thus, a targeted lytic reaction, such as those catalysed by enzymes, is essential to maintain DNA integrity. It was proven here that despite morphological or chemical changes that might occur in the bacterial cell wall during FFPE processing, enzymes in the Metapolyzyme mix are still effective at targeting and lysing FFPE bacteria (Figure 5), and that lysis with Metapolyzyme led to more uniform DNA recovery that more closely resembled the bacterial contents of a mock FFPE community. This is supported by evidence in archaic samples, where lysis with Metapolyzyme was found to yield a community composition that resembles that obtained by bead beating [43].

In addition to the two previously discussed sample treatments, there are other parameters to take into account when working with FFPE tissue. *(i)* FFPE tissues are paraffin embedded, dehydrated and carry traces of organic solutions (formalin, ethanol, and xylene) that can be toxic for enzymes. *(ii)* In FFPE samples, the tissues are tightly packed and hardened. This will limit access of enzymes to the target cells. *(iii)* All these processes require rapid changes of solution that establish the reaction

conditions required for the enzymes or active principles to take place. Therefore, integration of this protocol without the significant loss of bacterial cells / DNA is also a parameter to consider. For example, it has been noted that exposing G- bacteria to centrifugation speeds higher than 6,000-x g can damage the bacterial envelop, which will translate into the loss of bacterial DNA during processing [69]. Thus, repeatedly centrifuging samples at high speed for solution exchange could lead to bacterial cell wall damage, before treatment with DNase. Here the integration of the protocol was achieved by using sterile 0.2 µm CA filtering columns, which allowed a rapid flow rate with the lowest retention or losses. In addition, the inclusion of rehydration steps to the protocol was found to significantly increase amplifiable DNA yields. Lastly, a tissue dissociation step with the lowest off-target effects was adopted (Proteinase K).

During the assessment of the host depletion strategy, an off-target effect was observed in a qualitative analysis. This is unsurprising as most if not all host depletion strategies report some off target effects on bacteria [6]. To fully explore the effects that this would have on downstream sequence analysis, paired protoblock samples treated with (IHP) and without (IHN) host depletion were analysed by 16S sequencing and compared to the gold standard Qiagen QIAMP FFPE kit (Q). The results from this analysis indicate, that while there might be a loss of Gram negative bacteria, this does not significantly affect the outputs of 16S sequencing. This is supported by qPCR evidence showing higher recovery of the 16S gene by QUBIT and qPCR, in addition to an increase in the bacterial to host DNA ratio (sFigures 4). In addition to a better cross taxa representation of bacterial DNA recovered by qPCR (sFigure 5).

Furthermore, in this analysis the QIAGEN method, which is not optimised for bacterial lysis, showed statistically significant deviation from the input proportions across all five bacterial species present in the Protoblock. The IHN method showed improvement on the Q method, with the IHP method being the best performing approach in this instance. This improvement in performance is related to incorporation of a host depletion step, since it is the only variable tested here. It can be hypothesised that this may be due to (1) a reduction of a good portion of the contaminants (as shown in Figure 7) that improves the ratio of bacteria present in the samples being sequenced and (2) the reduction of mammalian DNA positively affects the PCR reaction, by improving the access to target sequences.

This was further explored in mouse FFPE faecal samples were exposed or not to a combined treatment with tissue dissociation and host depletion. Based on the evidence from the Protoblock-based comparison of the three methods, the expectation would be for the DT-P (in house with host depletion and tissue dissociation) and DT-N (in house without host depletion and tissue dissociation) to cluster together on a PcOA. However, in this instance, it was the DT-N and Q methods that clustered, showing no statistically significant difference in terms of their Bray-Curtis dissimilarity. Both are significantly different to the samples processed using the DT-P method. Subsequent spearman correlation of the dominant bacterial families identified across the samples with the PC1 and PC2 axis reveals that this separation on the PcOA plot is driven by Gram status. Gram positive bacteria correlate significantly with the direction of the DT-P samples, and Gram negative samples correlate significantly with the two other groups (Figure 9). These findings are corroborated by results in Figure 10. Altogether, these results confirm a significant loss of G- bacteria after the combined treatment with tissue dissociation and host depletion strategies, indicating that Proteinase K debilitates the OM of G- bacteria, exposing the phospholipid bilayer, which can be then accessed by Saponin, leading to G- bacteria loss. However, this is a necessary step in processing tissues, and thus a further optimisation of this step is necessary. This could be addressed by incorporating a short decrosslinking step that will allow tissue dissociation enzymes to be more effective, leading to a reduction on incubation times or enzyme units used in the reaction. This could lead to less off-target effects in G- bacteria.

By a process of elimination, the best net performing method in this instance appears to be the DT-N method. The DT-P method shows significantly increased Gram positive bacterial family abundance such as *Lactobacillaceae* and *Coriobacteriaceae* when compared with the Qiagen method; conversely the Qiagen method shows significantly more Gram negative bacteria such as *Prevotellaceae* and *Bacteroidaceae*. The DT-N method shows significantly more Gram negative bacterial families vs DT-P (Figure 10c), and significantly more Gram positive families such as *Coriobacteriaceae* vs Q, with no families significantly reduced in abundance vs either group. Thus confirming that the tissue dissociation strategy needs to be optimised.

Despite mayor efforts on maintaining an aseptic technique, there are still numerous potential sources of contamination, ranging from the wax used to embed samples, through all the DNA purification solutions and enzymes, which are unsuitable for sterilisation or could not be gamma irradiated at our facilities. Thus, it is unsurprising that there was a considerable amount of contamination present in the samples. The biomass in the Protoblock and Mouse Faeces samples is sufficient to ensure that the majority of the reads are of sample origin according to the SourceTracker algorithm, but the FFPE breast samples appeared to consist almost entirely of bacterial reads attributed to one or more of the negative controls. The SourceTracker output in Figure 6b indicates that all contamination is attributable to three negative control samples, namely the Wax control, taken from the edges of the blocks of patient samples, the "In House method" negative control, and the non-bacterial control, which is an empty Protoblock FFPE processed at our facilities. The first two negative controls are dominated by the genera *Stenotrophomonas, Pseudomonas* and *Clostridium*, all of which count among the most abundant genera in the dataset. The presence of both high and low abundance environmental contaminants presents a problem for most bioinformatic contamination removal methods, and highlights the value of using both positive and negative controls to assist in contamination removal [70]. In this instance, we are provided with a much clearer picture of the contamination induced during the process by the use of the Protoblock in conjunction with negative controls. Figure 6 also provided us with evidence of a phenomenon that is gaining more attention in Microbiome research, cross contamination, originating within the pool of samples[71]. This phenomenon is known to affect lower biomass samples, and can be clearly seen in the non-bacterial control where five of the common bacterial families across the dataset also appear in the negative controls. This is particularly dangerous when undertaking established, but conservative contamination removal by subtraction approaches.

Validation in FFPE breast tissue Non-tract biopsies are notoriously low in microbial biomass [72], a fact that is further compounded in analysis of FFPE biopsies by the fact that the formalin fixation process accounts for a log fold reduction in the quantity of recoverable DNA [73]. These challenges clearly manifest in the comparison of paired fresh and FFPE breast samples. Once the major contaminant

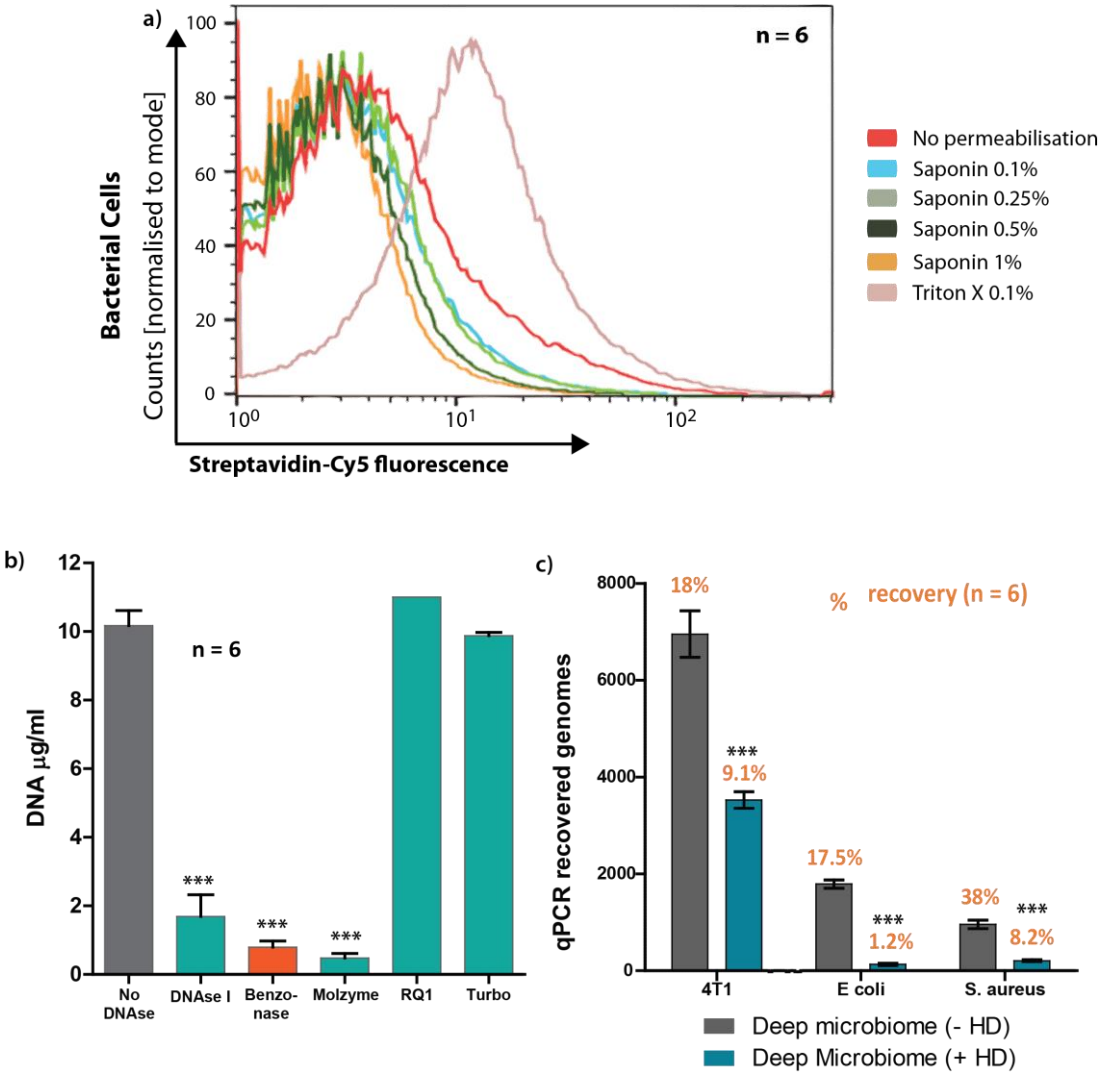
ASV's and those suspected of aligning to the human genome are removed, the FFPE breast samples are still dominated by known contaminant families, seen in the negative controls in Figure 6. Encouragingly, there are some common families to both the FFPE breast samples shown in Figure 11b and the fresh frozen breast sample shown in Figure 11c. As mentioned in the results, manual removal of the *Pseudomonadaceae* and *Xanthomonadaceae* families reveals a sample composition plot where 24.6% of the total bacterial abundance in FFPE breast tissue is accounted for by the bacterial families also present in the fresh frozen breast samples (Figure 11a).

The reason for Figure 11d is that it is a crude retrospective imitation of a potential improvement to make this method a viable option for low biomass FFPE studies. With the main contaminants inherent to the In House FFPE protocol now identified, these can be biologically removed from the sample by blocking their amplification from the 16S PCR pool. Numerous methods have been developed to achieve an asymmetric PCR reaction that will favour the amplification of certain target regions and avoid the amplification of other, which have been used extensively for SNP detection or to reduce off-target capture during sequencing library enrichment. This is achieved by: (1) Blocking extension with DNA probe/oligo that has high affinity towards a specific DNA sequence (on either DNA strand) that includes a 3' end (i.e. phosphate, inverted dNTP). (2) Inhibiting primer annealing with a homologous peptide nucleic acid (PMA) or locked nucleic acids (LNAs), which have increased thermal or base stacking stability, respectively and will inhibit PCR [74-76].

CONCLUSION

Strategies for the unbiased treatment of FFPE samples for microbiome analysis are presented in this work, as summarised in Figure 1. Each step validated by flow cytometry, qPCR and 16S sequencing on mock bacterial communities, murine models and human breast tissue samples. The results shown here confirm that most of these strategies would have a positive effect in the treatment for microbiome analysis. However, key areas that need to be addressed are the optimisation of a tissue dissociation strategy that does not lead to G- bacterial loss and the biological decontamination of samples previous to the analysis. Alternatives to achieve this are suggested.

SUPPLEMENTARY MATERIAL

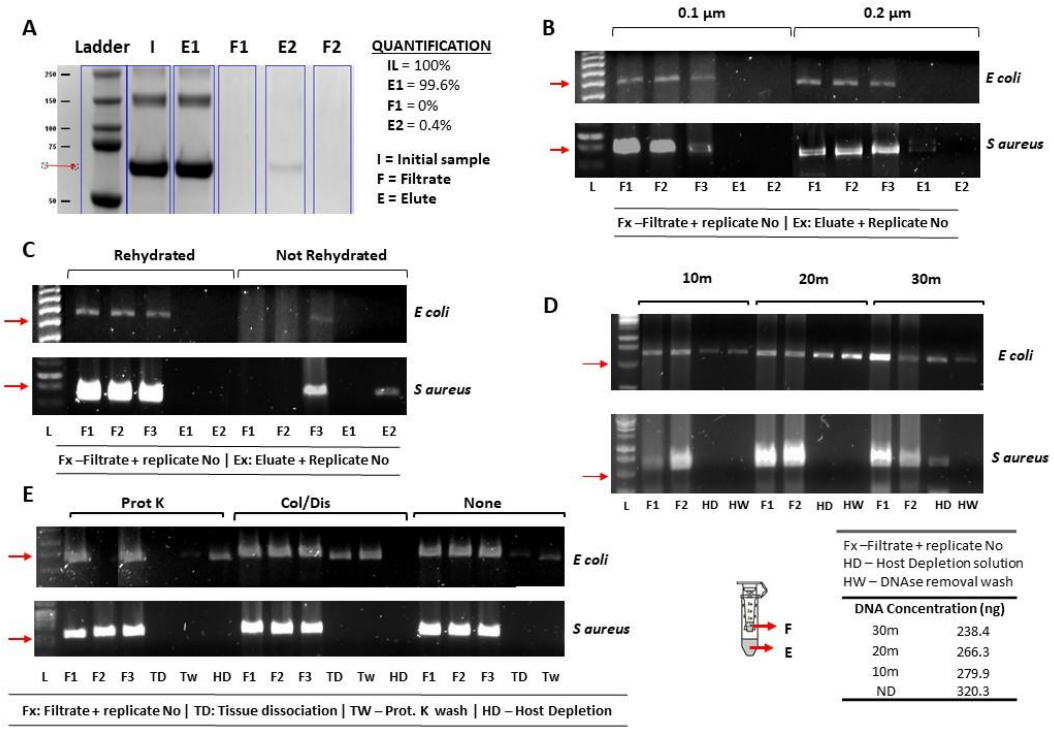


Supplementary Figure 1. Optimising host depletion.

(a) *<u>Saponin titration.</u> Histogram showing fluorescence intensity for Cy5. As a marker for protein internalisation, hence, membrane permeabilisation. E. coli cells were permeabilised with increasing concentrations of Saponin (n = 6, for each line). Unlike Triton-X (Pink), Saponin treated E. coli cells showed no increase in fluorescence intensity even after treatment with high (1%) saponin concentrations.*

(b) *<u>DNAse screen.</u> 5 commercially available DNAses were tested for their capacity of depleting DNA from 5 x 10^6 FF 4T1 cells in 20 min at 37°C. In bar plot is the resulting DNA yield obtained after DNA purification. From here, Benzonase was take as the most cost-effective strategy.*

(c) *<u>Trial with Molysis Host Depletion strategy.</u> Slides* from *FFPE blocks loaded with 4T1, E. coli and S. aureus cells were treated with the protocol, including (blue) or excluding (grey) DNA depletion. A decrease in quantity of recoverable DNA was observed in the 3 cell types.*



Supplementary Figure 2. Integration of the protocol

For all figures: ***L =*** *ladder.* ***Red arrows =*** *indicate the 500 bp mark in ladder. PCR targets a ~500 bp DNA fragment.* ***Eluates =*** *a total of 300 µl were DNA purified, 100 µl were aliquoted from each experimental replicate ($\bar{x}$ n = 3). Filtrates correspond to final filtrate, obtained at the end of the protocol. (A-B) were performed with an E. coli cell suspension. (C-E) were performed in slides from protoblocks with 1 x 10 ^{8} E. coli and S. aureus cells fixed for 48 h.*

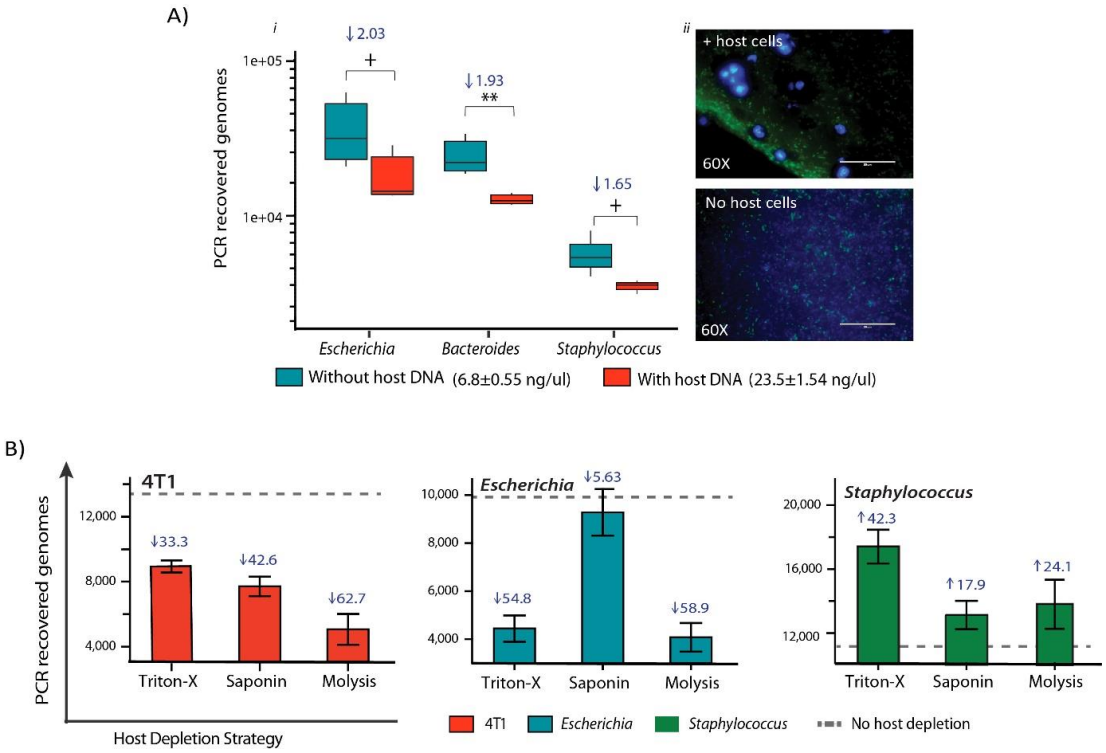
(A) *<u>Are proteins removed by filtering columns?</u> Protein gel confirming the removal of BSA (64 KDa) from a cell suspension.*

(B) *<u>Are bacteria lost when using a 0.2 µm filter?</u> Agarose gel of eluate and filtrates performed on a bacterial cell suspension washed twice with PBS. No DNA is seen for the eluates.*

(C) *<u>Does rehydrating the samples have any effect?</u> DNA gel showing eluate/filtrates after treatments. Only for rehydrated samples was there a band in all elutes. E1 = Host depletion, E2 = Wash after host depletion.*

(D) *<u>Is bacterial DNA lost by HD strategy? What is the optimal incubation time?</u> Gel showing that there is a loss of E. coli DNA during the HD strategy and this increase with longer incubations. Shorter incubation times (10m) reduces the loss of G - bacteria, but also reduces the quantity of host DNA depleted seen in the table below gels. Quantity as the difference from DNA concentration measured for non-depleted and the average DNA concentration (n = 3) obtained after host depletion for each time point.*

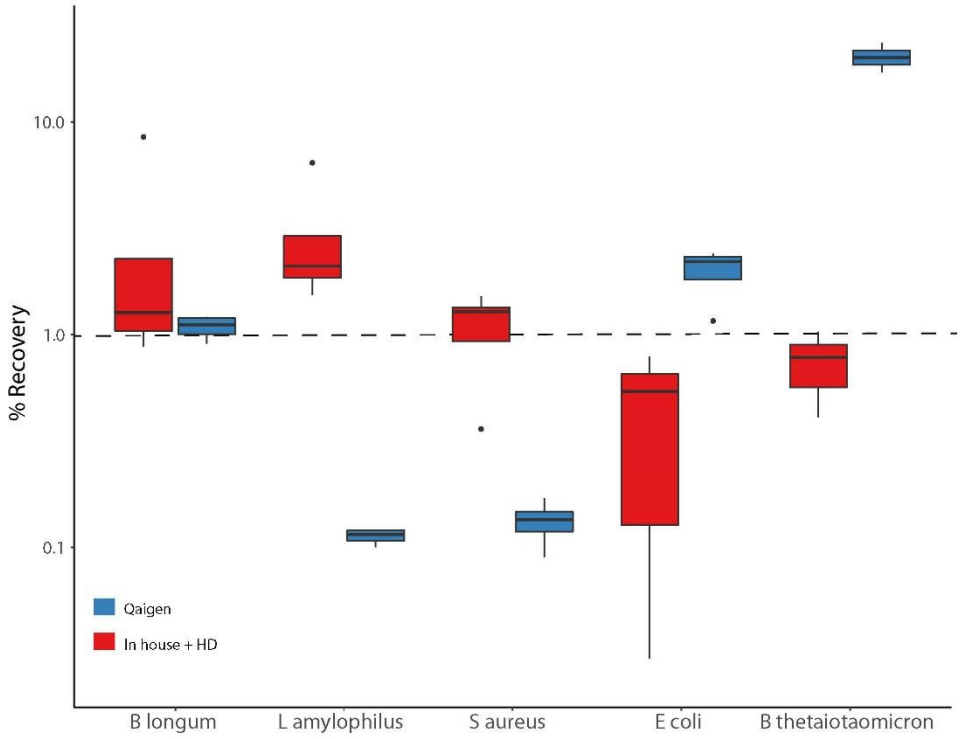
(E) *<u>Do tissue dissociation strategies affect bacteria? Which is the most adequate tissue dissociation strategy? Gel</u> showing that there is loss of G- bacteria during tissue dissociation. This is less pronounced for Proteinase K, which was deemed the most adequate.*



Supplementary Figure 3. Investigating Host DNA influence and host DNA depletion

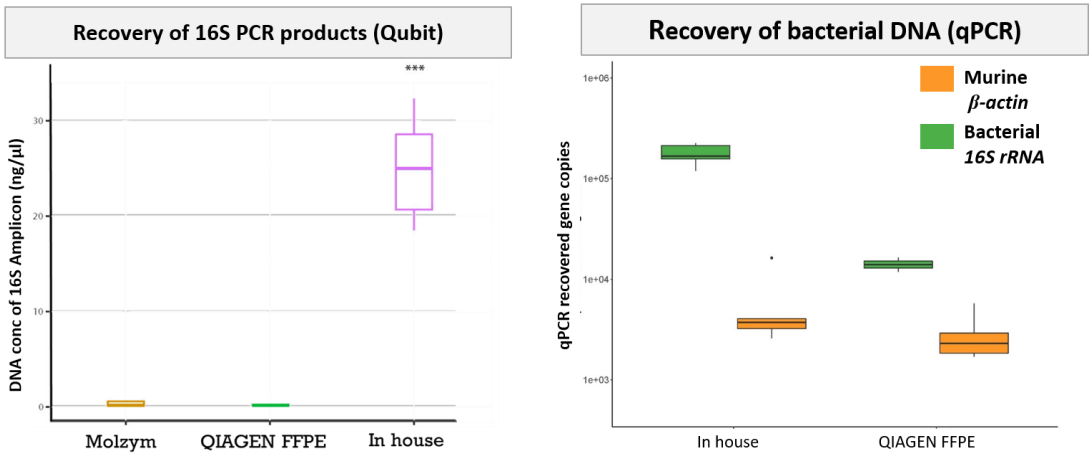
(A) Measuring bias introduced by host DNA. i) *Box plot comparing DNA recovery of bacteria in Protoblocks loaded with (cyan) and without 4T1 cells (orange). Quantitative PCR recovery was normalised to a sample input of 10^6 cells. For each box, n = 6. Protoblocks without 4T1 cells had a higher recovery of all bacteria taxa. Difference of means between tests was measured using a Wilcoxon Signed Rank test, for all bacterial taxa.* ***ii)*** *Immunofluorescence microscopy images of Protoblocks with and without mammalian cells, stained with α-E.coli (green) and DAPI (Blue) for 4T1 cells.*

(B) Testing host DNA depletion strategies*. DNA recovery of 4T1 cells (orange), Escherichia (cyan) and Staphylococcus (green) after 10 min treatment with either Triton-X (0.1%), Saponin (0.1%) or Molysis CM buffer. For each bar, n = 3. % increase or decrease in recovery from untreated is shown above each bar. Dotted lines indicate the PCR recovery of samples without host depletion. (In all cases p = . <0.1, * < 0.05, ** < 0.01, *** < 0.001)*



Supplementary Figure 4. QPCR percent recovery of bacterial DNA from protoblocks. In red In house method + Host Depletion.

In blue: QIAGEN FFPE kit. The qPCR reaction was set in a multiplex format, using primers and probes described in methods. A recovery of 1% is a normal recovery, lower % recovery is considered a reduction and higher % an overestimation.



Supplementary Figure 5. Measuring the recovery of 16S amplicons.

a) *DNA concentration of amplicons recovered after 16S rRNA gene PCR for protoblocks processed with QIAGEN FFPE DNA kit, Molzym, and the in house method.* ***b)*** *QPCR recovery of B-actin or the 16S gene from protoblocks processed with the in house method or QIAGEN FFPE DNA. It is clear from this figure that with the in house method a higher recovery of*

bacterial DNA (16S gene) is achieved, at the same time an improvement in the bacterial to host DNA ratio is achieved.

Supplementary Table 1. Bacterial load of protoblocks used for 16S sequencing

	Counts in microscope / volume measured for each type of FFPE block [single strain to mixed strain]			Calculations for DNA purified from blocks			
Cell type	Microscope Counts in block	Cells/µl in mixed block	Cells in 15 µm slide	Cells DNA extraction	Genomes in elution	Cells in 16S PCR	Ratio
4T1	2.20E+07	8.85E+04	1.06E+06	1.28E+07	2.55E+05	**3.83E+06**	
Escherichia coli	3.10E+07	1.25E+05	1.46E+06	1.75E+07	3.50E+05	**5.25E+06**	0.17
S. aureus	9.01E+06	3.63E+04	4.22E+05	5.06E+06	1.01E+05	**1.52E+06**	0.05
B. longum	8.50E+06	3.43E+04	4.01E+05	4.81E+06	9.62E+04	**1.44E+06**	0.05
L. amylophilus	8.12E+07	3.28E+05	3.74E+06	4.48E+07	8.97E+05	**1.35E+07**	0.43
B. thetaiotao	5.82E+07	2.34E+05	2.71E+06	3.26E+07	6.51E+05	**9.77E+06**	0.31

REFERENCES

1. Bøifot, K.O., et al., *Performance evaluation of a new custom, multi-component DNA isolation method optimized for use in shotgun metagenomic sequencing-based aerosol microbiome research.* bioRxiv, 2019: p. 744334.

2. Riquelme, E., et al., *Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes.* Cell, 2019. **178**(4): p. 795-806.e12.

3. Clooney, A.G., et al., *Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis.* PLoS One, 2016. **11**(2): p. e0148028.

4. Salter, S.J., et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.* BMC Biology, 2014. **12**(1): p. 87.

5. Chen, L., et al., *DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification.* Science, 2017. **355**(6326): p. 752.

6. Marotz, C.A., et al., *Improving saliva shotgun metagenomics by chemical host DNA depletion.* Microbiome, 2018. **6**(1): p. 42.

7. Pereira-Marques, J., et al., *Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis.* Frontiers in Microbiology, 2019. **10**(1277).

8. Fricker, A.M., D. Podlesny, and W.F. Fricke, *What is new and relevant for sequencing-based microbiome research? A mini-review.* Journal of Advanced Research, 2019. **19**: p. 105-112.

9. Wu, J.Y., et al., *Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method.* BMC Microbiol, 2010. **10**: p. 255.

10. le Maire, M., P. Champeil, and J.V. Møller, *Interaction of membrane proteins and lipids with solubilizing detergents.* Biochimica et Biophysica Acta (BBA) - Biomembranes, 2000. **1508**(1): p. 86-111.

11. van Meer, G., D.R. Voelker, and G.W. Feigenson, *Membrane lipids: where they are and how they behave.* Nature reviews. Molecular cell biology, 2008. **9**(2): p. 112-124.

12. Miller, S.I. and N.R. Salama, *The gram-negative bacterial periplasm: Size matters.* PLoS biology, 2018. **16**(1): p. e2004935-e2004935.

13. Silhavy, T.J., D. Kahne, and S. Walker, *The bacterial cell envelope.* Cold Spring Harbor perspectives in biology, 2010. **2**(5): p. a000414-a000414.

14. May, K.L. and M. Grabowicz, *The bacterial outer membrane is an evolving antibiotic barrier.* Proceedings of the National Academy of Sciences, 2018. **115**(36): p. 8852-8854.

15. Harder, T., et al., *Lipid Domain Structure of the Plasma Membrane Revealed by Patching of Membrane Components.* The Journal of Cell Biology, 1998. **141**(4): p. 929-942.

16. Nelson, M.T., et al., *Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles.* Cell Reports, 2019. **26**(8): p. 2227-2240.e5.

17. Oechslin, C.P., et al., *Limited Correlation of Shotgun Metagenomics Following Host Depletion and Routine Diagnostics for Viruses and Bacteria in Low Concentrated Surrogate and Clinical Samples.* Frontiers in Cellular and Infection Microbiology, 2018. **8**(375).

18. Horz, H.-P., et al., *Selective isolation of bacterial DNA from human clinical specimens.* Journal of Microbiological Methods, 2008. **72**(1): p. 98-102.

19. Thoendel, M., et al., *Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing.* Journal of Microbiological Methods, 2016. **127**: p. 141-145.

20. Sachs, J.N. and T.B. Woolf, *Understanding the Hofmeister Effect in Interactions between Chaotropic Anions and Lipid Bilayers: Molecular Dynamics Simulations.* Journal of the American Chemical Society, 2003. **125**(29): p. 8742-8743.

21. Sawyer, W.H. and J. Puckridge, *The dissociation of proteins by chaotropic salts.* J Biol Chem, 1973. **248**(24): p. 8429-33.

22. Metz, B., et al., *Identification of Formaldehyde-induced Modifications in Proteins: REACTIONS WITH MODEL PEPTIDES.* Journal of Biological Chemistry, 2004. **279**(8): p. 6235-6243.

23. Feehery, G.R., et al., *A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA.* PLOS ONE, 2013. **8**(10): p. e76096.

24. Mattei, B., et al., *Membrane permeabilization induced by Triton X-100: The role of membrane phase state and edge tension.* Chemistry and Physics of Lipids, 2017. **202**: p. 28-37.

25. Niklas, J., et al., *Selective permeabilization for the high-throughput measurement of compartmented enzyme activities in mammalian cells.* Analytical Biochemistry, 2011. **416**(2): p. 218-227.

26. Stochaj, U., et al., *Nuclear envelopes show cell-type specific sensitivity for the permeabilization with digitonin.* Protocol Exchange, 2010.

27. Seeman, P., D. Cheng, and G.H. Iles, *Structure of membrane holes in osmotic and saponin hemolysis.* The Journal of cell biology, 1973. **56**(2): p. 519-527.

28. Amidzadeh, Z., et al., *Assessment of different permeabilization methods of minimizing damage to the adherent cells for detection of intracellular RNA by flow cytometry.* Avicenna journal of medical biotechnology, 2014. **6**(1): p. 38-46.

29. Meunier, E. and P. Broz, *Quantification of Cytosolic vs. Vacuolar Salmonella in Primary Macrophages by Differential Permeabilization.* Journal of visualized experiments : JoVE, 2015(101): p. e52960-e52960.

30. Marathe, S.A., et al., *Differential modulation of intracellular survival of cytosolic and vacuolar pathogens by curcumin.* Antimicrobial agents and chemotherapy, 2012. **56**(11): p. 5555-5567.

31. Teng, F., et al., *Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling.* Scientific Reports, 2018. **8**(1): p. 16321.

32. Abusleme, L., et al., *Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing.* Journal of Oral Microbiology, 2014. **6**(1): p. 23990.

33. Fiedorová, K., et al., *The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery.* Frontiers in microbiology, 2019. **10**: p. 821-821.

34. Knudsen, B.E., et al., *Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition.* mSystems, 2016. **1**(5): p. e00095-16.

35. Methé, B.A., et al., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.

36. Gill, C., et al., *Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota.* PLOS ONE, 2016. **11**(9): p. e0163148.

37. Bag, S., et al., *An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples.* Scientific Reports, 2016. **6**(1): p. 26775.

38. Einaga, N., et al., *Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.* PLOS ONE, 2017. **12**(5): p. e0176280.

39. Zhang, P., et al., *The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies.* International Journal of Genomics, 2017. **2017**: p. 9.

40. Yuan, S., et al., *Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome.* PLOS ONE, 2012. **7**(3): p. e33865.

41. Liesack, W., H. Weyland, and E. Stackebrandt, *Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria.* Microbial Ecology, 1991. **21**(1): p. 191-198.

42. Tighe, S., et al., *Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP).* Journal of biomolecular techniques : JBT, 2017. **28**(1): p. 31-39.

43. Zaikova, E., et al., *Antarctic Relic Microbial Mat Community Revealed by Metagenomics and Metatranscriptomics.* Frontiers in Ecology and Evolution, 2019. **7**(1).

44. Cronin, M., et al., *High resolution in vivo bioluminescent imaging for the study of bacterial tumour targeting.* PLoS One, 2012. **7**(1): p. e30940.

45. Blum-Oehler, G., et al., *Development of strain-specific PCR reactions for the detection of the probiotic Escherichia coli strain Nissle 1917 in fecal samples.* Res Microbiol, 2003. **154**(1): p. 59-66.

46. Madison, B.M. and V.S. Baselski, *Rapid identification of Staphylococcus aureus in blood cultures by thermonuclease testing.* Journal of clinical microbiology, 1983. **18**(3): p. 722-724.

47. Blaiotta, G., et al., *Lactobacillus strain diversity based on partial hsp60 gene sequences and design of PCR-restriction fragment length polymorphism assays for species identification and differentiation.* Applied and environmental microbiology, 2008. **74**(1): p. 208-215.

48. Teng, L.J., et al., *PCR assay for species-specific identification of Bacteroides thetaiotaomicron.* J Clin Microbiol, 2000. **38**(4): p. 1672-5.

49. Altmann, F., et al., *Genome Analysis and Characterisation of the Exopolysaccharide Produced by Bifidobacterium longum subsp. longum 35624.* PLoS One, 2016. **11**(9): p. e0162983.

50. Biologicals, N., *Immunohistochemistry (IHC) Handbook* BIOTECHNE, Editor. 2017.

51. Suck, D., C. Oefner, and W. Kabsch, *Three-dimensional structure of bovine pancreatic DNase I at 2.5 A resolution.* The EMBO journal, 1984. **3**(10): p. 2423-2430.

52. Kuzuya, A., et al., *Precisely programmed and robust 2D streptavidin nanoarrays by using periodical nanometer-scale wells embedded in DNA origami assembly.* Chembiochem, 2009. **10**(11): p. 1811-5.

53. Chaiet, L. and F.J. Wolf, *The properties of streptavidin, a biotin-binding protein produced by Streptomycetes.* Archives of Biochemistry and Biophysics, 1964. **106**: p. 1-5.

54. Zempleni, J., S.S.K. Wijeratne, and Y.I. Hassan, *Biotin.* BioFactors (Oxford, England), 2009. **35**(1): p. 36-46.

55. Liu, F., J.Z.H. Zhang, and Y. Mei, *The origin of the cooperativity in the streptavidin-biotin system: A computational investigation through molecular dynamics simulations.* Scientific Reports, 2016. **6**(1): p. 27190.

56. Schiapparelli, L.M., et al., *Direct detection of biotinylated proteins by mass spectrometry.* Journal of proteome research, 2014. **13**(9): p. 3966-3978.

57. Bailey, L.M., et al., *Artifactual detection of biotin on histones by streptavidin.* Analytical Biochemistry, 2008. **373**(1): p. 71-77.

58. Bramwell, M.E., *Characterization of biotinylated proteins in mammalian cells using 125I-streptavidin.* J Biochem Biophys Methods, 1987. **15**(3-4): p. 125-32.

59. Dundas, C.M., D. Demonte, and S. Park, *Streptavidin–biotin technology: improvements and innovations in chemical and biological applications.* Applied Microbiology and Biotechnology, 2013. **97**(21): p. 9343-9353.

60. Sengüven, B., et al., *Comparison of methods for the extraction of DNA from formalin-fixed, paraffin-embedded archival tissues.* International journal of medical sciences, 2014. **11**(5): p. 494-499.

61. Sigma, M., *Benzonase® endonuclease*.

62. Hasan, M.R., et al., *Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing.* Journal of clinical microbiology, 2016. **54**(4): p. 919-927.

63. Charalampous, T., et al., *Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection.* Nat Biotechnol, 2019. **37**(7): p. 783-792.

64. Cacciatore, S., et al., *Metabolic Profiling in Formalin-Fixed and Paraffin-Embedded Prostate Cancer Tissues.* Molecular cancer research : MCR, 2017. **15**(4): p. 439-447.

65. von Wintzingerode, F., U.B. Gobel, and E. Stackebrandt, *Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis.* FEMS Microbiol Rev, 1997. **21**(3): p. 213-29.

66. Liesack, W., H. Weyland, and E. Stackebrandt, *Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria.* Microb Ecol, 1991. **21**(1): p. 191-8.

67. Yuan, S., et al., *Evaluation of methods for the extraction and purification of DNA from the human microbiome.* PloS one, 2012. **7**(3): p. e33865-e33865.

68. Natarajan, V.P., et al., *A Modified SDS-Based DNA Extraction Method for High Quality Environmental DNA from Seafloor Environments.* Frontiers in Microbiology, 2016. **7**(986).

69. Peterson, B.W., et al., *Bacterial cell surface damage due to centrifugal compaction.* Applied and environmental microbiology, 2012. **78**(1): p. 120-125.

70. Eisenhofer, R., et al., *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations.* Trends Microbiol, 2019. **27**(2): p. 105-117.

71. Minich, J.J., et al., *Quantifying and Understanding Well-to-Well Contamination in Microbiome Research.* mSystems, 2019. **4**(4): p. e00186-19.

72. Jervis-Bardy, J., et al., *Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data.* Microbiome, 2015. **3**: p. 19-19.

73. Hykin, S.M., K. Bi, and J.A. McGuire, *Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing.* PloS one, 2015. **10**(10): p. e0141579-e0141579.

74. Vestheim, H., B.E. Deagle, and S.N. Jarman, *Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR.* Methods Mol Biol, 2011. **687**: p. 265-74.

75. Bender, M., et al., *Use of a PNA probe to block DNA-mediated PCR product formation in prokaryotic RT-PCR.* Biotechniques, 2007. **42**(5): p. 609-10, 612-4.

76. Wang, H., et al., *Allele-specific, non-extendable primer blocker PCR (AS-NEPB-PCR) for DNA mutation detection in cancer.* J Mol Diagn, 2013. **15**(1): p. 62-9.

CHAPTER 5:

Discussion

In this manuscript are presented the foundations for to pursuing microbiome analysis on FFPE samples. FFPE samples represent a huge potential repertoire of material for microbiome studies yet to be accessed. Finding of bacterial profiling in FFPE samples have encouraged the use of this samples as source material by providing evidence that bacteria remain in tissues after FFPE samples and that the communities are still somehow representative of what has been previously observed in fresh samples. However, the lack of evidence describing the state of bacteria in FFPE tissues is a mayor limitation, which might even lead to doubt findings from projects using them.

Pursuing microbiome studies in human tissues has been proven challenging in non-fixed samples, as many obstacles such as low bacterial biomass, high human DNA background, high levels of contamination from surgical processing and the lack of a standardised protocol for these sample-type constrain the scope and extent of the analysis. Now, in FFPE samples, it can be assumed that these would be more accentuated by the unsterile nature of the FFPE process and the detrimental effects that formalin exerts on biomolecules. Nevertheless, giving its huge potential as resource, exploring this sample-type and optimising methods for their use as source material could be a corner-stone for microbiome research. This will allow for retrospective research, provide access to tissues from inaccessible body sites and study the microbiome of numerous diseases for which FFPE samples have been catalogued. It will also facilitate current clinical studies, by avoiding interference during surgical processes.

Now, to address the gap of knowledge of the state of bacteria in FFPE samples it was first required to bridge several research areas: histology, microbiology, cancer biology and genomics. State of the art knowledge was gathered to raise the questions needed testing and to challenge assumptions that could be carried in translation from one field to the other. For example: Is formalin-fixation the same in bacteria then in human? Is DNA damage in bacteria the same as in human DNA? Are bacterial cells permeabilised/lysed by processing and hence do not necessitate further lysis? Which are the best models to study these?

The first issue that needed to be addressed was a model suitable to describe FFPE bacteria. The first attempts were done in mouse tissue loaded with bacterial cells,

however when using this model sample to sample variability was too high, impeding the attribution of results to experimental treatments. This could be due to differences in the distribution of bacterial cells within the tissue and the morphological differences of each layer of tissue included in each sample. Therefore, a simplified model that allowed for a reduction in variability and more uniform results was sought and developed by embedding fixed bacteria with or without mammalian cells in an agar mould that was processed using standard histology procedures. This model exhibited uniform results and allowed for a better description of molecular and cellular changes that can be attributed exclusively to the fixation process, and enabled the measurement of treatment effects that could lead to improvements in the yields and quality of DNA.

This allowed for a description of the state of FFPE DNA in bacteria. It was confirmed here that bacterial FFPE DNA is highly fragmented, crosslinked and bears sequence artefacts that reflect oxidative damage, uracil and methylation. The information gathered here and that found for human FFPE samples, leads to hypothesize that these damages are more pervasive in bacteria. This, given that FFPE blocks studied here were not stored for more than one year and that damage in human DNA to the extent found here is usually found in blocks stored for more than three years. Among differences found between bacterial and human DNA damage, is a higher rate of oxidative damage than reported for humans, with uracil lesions being less pervasive. However, the ratio at which these occur in human FFPE DNA is still a topic of controversy, as different ratios are found on different samples. As an off-topic conclusion, it is suggested that an evaluation of human FFPE DNA damage is performed in a simplified model, such as the protoblock (but with only mammalian cells). This will allow for an assessment without conflicting results that derive from inter-sample variability.

The information gathered by the evaluation of DNA integrity in FFPE bacterial allowed the development of strategies that reduce this damage or improves its quality. This was achieve by reducing DNA denaturation and its associated breaking, with a combination of decrosslinking at lower temperatures in the presence of chaotrophic agents and its repair with the Base Excision Repair pathway targeting oxidative DNA damage.

In addition, the state of the bacterial envelop/wall was investigated. Results from this study suggest that the FFPE bacterial cell envelop and wall are not permeable to large molecules, such as DNAses or Proteinase k. Thus, allowing for the development of differential lysis strategies that allow for host depletion. Here, a cholesterol targeting detergent, saponin, proved effective for differential lysis. Furthermore, this also informed of the necessity to incorporate a bacterial lysis strategy, which was proven effective here with a mix of bacterial lytic enzymes, Metapolyzyme, followed by a prolonged protein digestion with Proteinase K.

Evidence gathered by flow cytometry, qPCR and 16S analysis after treatment with the combined strategies for host depletion, bacterial lysis, protein digestion, decrosslinking and DNA repair indicate that the quantity and quality of bacterial FFPE DNA yielded outperforms those gold-standard kits for either FFPE human DNA. Furthermore, the bacterial community composition obtained is more resembling to that of the input, again outperforming the FFPE human DNA kit or non-fixed microbiome kit. Therefore, these strategies can be readily optimised for adoption in the microbiome analysis workflow, in FFPE samples.

On the other hand, it is suggested from the results obtained that new strategies for tissue dissociation, should be tested, as the strategies tested here debilitate the gram-negative bacterial envelop, which is detrimental to the downstream workflow. In this study, no other strategies that would not be detrimental for bacteria, were found, thus tested, however, this search was not done exhaustively. It can be hypothesised that with a mild, short period decrosslinking incubation before tissue dissociation would allow a faster digestion of tissue fibres, thus lower enzyme concentrations or incubation times would be used. Thus, limiting off-target activity in bacteria.

Besides, it was decided here to use filtering columns to allow rapid solution exchange and reduce the probabilities of damaging the bacterial cell walls or losing bacteria during washes by centrifugation. However, these proved more difficult to handle when testing them with tissue specimens, as they will clot and delay buffer exchange. It is advisable that when working with large tissue sections the tissue input is limited to 2 -3 slide. Otherwise, alternatives for carrying out this process should be sought.

Furthermore, despite made efforts in maintaining aseptic procedures along the process, a large amount of contamination in low-biomass samples obscured their analysis. These contaminants were traced to: (1) Reagent prepared in the lab: All solutions were prepared aseptically in a laminar hood, filter-sterilised, autoclaved (if allowed by the solution) or UV sterilised for 20 min. Despite all of these efforts, unless reagents are prepared in clean-room environment and gamma-irradiated, bacterial DNA will be present and present a mayor problem for low biomass samples. (2) FFPE processing: It was clear from the processing controls that some contaminant bacterial DNA was sourced from processing of FFPE samples. The identification of these contaminants is vital to proceed with FFPE-microbiome workflows, as they inform strategies for bioinformatic or biological removal. It was shown here that the bioinformatic decontamination of samples is in cases not sufficient to obtain a valid sequencing analysis. As such, it is important to include in the workflow strategies that allow for the biological removal of contaminating sequences from the PCR pool. This can be easily achieved through PCR enrichment methods, such as blocking DNA probe/oligos, peptide nucleic acid (PMA) or locked nucleic acids (LNAs) [1-3].

Lastly, there are variables in the tissue processing and storage conditions of FFPE samples that were not in the scope of this project and need to be addressed before adopting a protocol for microbiome analysis. While the models used here were representative of long fixation periods, as a mean of keeping a 'worse fixation scenario', due to time constraints in the duration of this project the maximum storage time assessed was 1 year. Therefore, a study characterisation the damage with age would be needed as a comparison for older FFPE samples.

Altogether, it is concluded from this work that to unlock the huge potential that FFPE specimens could provide to the field of microbiome research, it is essential that dedicated workflows designed for this sample type need to be in place. While sample prep was shown here as fundamental, this workflow should not be restricted to the sample-prep, and must include a robust QC system that allows for the screening of DNA quality in a sample and directs the workflow to either reject the sample, perform a DNA repair strategy or directs the amplification strategies. . In addition, a database for known FFPE derived contaminants should be in place to inform future potential strategies for their biological removal. Such workflow should include a dedicated 16S

sequencing workflow optimal for low-biomass FFPE samples, which might require longer PCR cycles, the use of shorter 16S sequences (e.g. V1-V2), or the use of more amplicon template for the sequencing runs. All sequencing workflow available to date are optimised for the analysis of high-biomass faecal samples, which differ significantly to FFPE specimens. As shown in this work, both sample types cannot be analysed with the same workflows.

REFERENCES

1. Vestheim, H., B.E. Deagle, and S.N. Jarman, *Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR.* Methods Mol Biol, 2011. **687**: p. 265-74.

2. Bender, M., et al., *Use of a PNA probe to block DNA-mediated PCR product formation in prokaryotic RT-PCR.* Biotechniques, 2007. **42**(5): p. 609-10, 612-4.

3. Wang, H., et al., *Allele-specific, non-extendable primer blocker PCR (AS-NEPB-PCR) for DNA mutation detection in cancer.* J Mol Diagn, 2013. **15**(1): p. 62-9.