

Title	Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions
Authors	Claesson, Marcus J.;Wang, Qiong;O'Sullivan, Orla;Greene-Diniz, Rachel;Cole, James R.;Ross, R. Paul;O'Toole, Paul W.
Publication date	2010
Original Citation	Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P. and O'Toole, P. W. (2010) 'Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions', Nucleic Acids Research, 38(22), e200 (13pp). doi: 10.1093/nar/gkq873
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq873 - 10.1093/nar/gkq873
Rights	© 2010, the Authors. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited. - http://creativecommons.org/licenses/by-nc/2.5
Download date	2025-07-05 02:34:35
Item downloaded from	https://hdl.handle.net/10468/5027



UCC

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Dinucleotide repeat polymorphism at the d10s179 locusSUPPLEMENTARY DATA

Filtering and trimming trials to improve read quality

To remove poor quality sequences, several methods of quality-filtering were tried on the raw reads in addition to the standard filtering protocol described in the main text. In one set of trials, reads were trimmed to fixed lengths of either 50 or 75 bases, including primers (labelled 50bp and 75bp filters, respectively). Trimmed reads with an average Q score below 20 were discarded, as were sequences with one or more ambiguous bases. In addition, trimmed reads were discarded if they contained more than 2 or 3 bases with Q score below 10, or more than 45 or 70 identical bases for the 50bp and 75bp filters, respectively. For two additional trials, reads were trimmed at the first base with a Q score below 10 or a Q score below 20 (labelled 10qv and 20qv filters, respectively). Thus, the two first filtering criteria resulted in fixed length reads, whereas that last two criteria in mixed length reads. For all filters, after removing primers, the paired reads were discarded if one end failed the quality filter or if the combined length of the paired reads was below 50 bases. For comparison purposes this minimum 50 base requirements was added to the standard filter results. The percentage of reads that passed these quality filters varied by region (Suppl. Fig. S1), and only small fractions of reads passed the 75bp and 20qv filters.

The resulting quality-filtered and trimmed reads were assigned to phyla using the RDP-classifier with a bootstrap cut-off of 50% and were subsequently compared to similar results using the standard filtering protocol (Suppl. Fig. S2). In general, the total percent classifiable reads at the phylum level increased with 75 bp filter, but decreased with the 10qv and 20qv filtering and trimming strategies. In addition, the overall taxonomic composition of the trimmed and filtered reads changed dramatically when compared with the standard filtering strategy, especially for the 75bp and 20qv filters. The effect was apparent in all amplicons but less prominent for the V4V5 amplicon. Only a small percentage of sequences changed classification after trimming (Table S1), indicating that the most likely reason for the observed changes in taxonomic composition was differential effects of filtering on the different phyla.

Comparative compositional microbiota analysis of duplicate samples

We compared microbiota compositions of four duplicate samples in order to see if these would group together in spite of not having identical profiles of relative phylum/genus abundances. Fresh faecal samples from four other elderly individuals were processed in the same way as the single analyzed sample, but with 40,000 V4 amplicons sequenced per sample. DNA from samples collected from subjects EM1, EM2 and EM3 were extracted only once, but with the 16S rRNA V4 sequence amplified twice using the same conditions (technical replicates). Subject EM4, on the other hand, had been sampled at two different time-points, one day apart (biological replicate). After sequencing on a 454 FLX machine, the reads were subjected to the same phylogenetic analysis as for the single sample (Suppl. Fig. S3). In addition, hierarchical UniFrac analysis (1) was also performed, following alignment of unique reads using Infernal (2) and tree-building using FastTree (3). Both weighted (including abundances of unique reads) and un-weighted (only including presence/absence of unique reads) UniFrac distances was performed to compare the samples using the phylogenetic information (Suppl. Fig. S4).

Even though the branching topologies differ somewhat between the two trees, all replicates group together, irrespective of the type of UniFrac analysis (weighted or un-weighted). Observations that inter-individual variation of human microbiomes are generally higher than intra-individual variation, be it technical or biological replicates, have been extensively reviewed by others (4). In contrast, the phylogenetic profiles (Suppl. Fig. S3) of the replicates are markedly different at both phylum and genus level; even at the coarsest phylum-level there are clear and noticeable differences between all replicates. This illustrates that finer resolution of microbiota (unique reads) can offer higher discriminatory power than at phylum or genus levels (see also ref (5)), and that classifications exclusively to the latter levels should not be done when comparing a multitude of samples. However, it is important to note that this type of clustering analysis is not possible when comparing different types of amplicons; UniFrac requires phylogenetic trees built from alignment of representative sequences (unique reads in this study) from the same 16S rRNA region.

References

1. Hamady, M., Lozupone, C. and Knight, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *Isme J*, **4**, 17-27.
2. Nawrocki, E.P. and Eddy, S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol*, **3**, e56.
3. Price, M.N., Huang, K.H., Alm, E.J. and Arkin, A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, **33**, 880-892.
4. Kuczynski, J., Costello, E.K., Nemergut, D.R., Zaneveld, J., Lauber, C.L., Knights, D., Koren, O., Fierer, N., Kelley, S.T., Ley, R.E. *et al.* (2010) Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol*, **11**, 210.
5. Lauber, C.L., Zhou, N., Gordon, J.I., Knight, R. and Fierer, N. (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol Lett*.

Table S1. Percent of reads passing the indicated filter that were assigned differently at phylum-level than when processed using the standard filter using the RDP-Classifer with a bootstrap cut-off of 50%.

Standard	V1V2	V2V3	V3V4	V4V5	V5V6	V7V8
50bp	0.85	0.11	2.48	0.29	0.60	1.18
75bp	0.01	0.03	0.06	0.07	0.11	0.27
10qv	0.26	0.13	0.39	0.07	0.32	0.71
20qv	0.43	0.03	1.53	0.10	0.29	0.69

Figure S1. Percentage of reads passing the various quality and trimming filters tested compared to the percentage passing the standard filtering procedure.

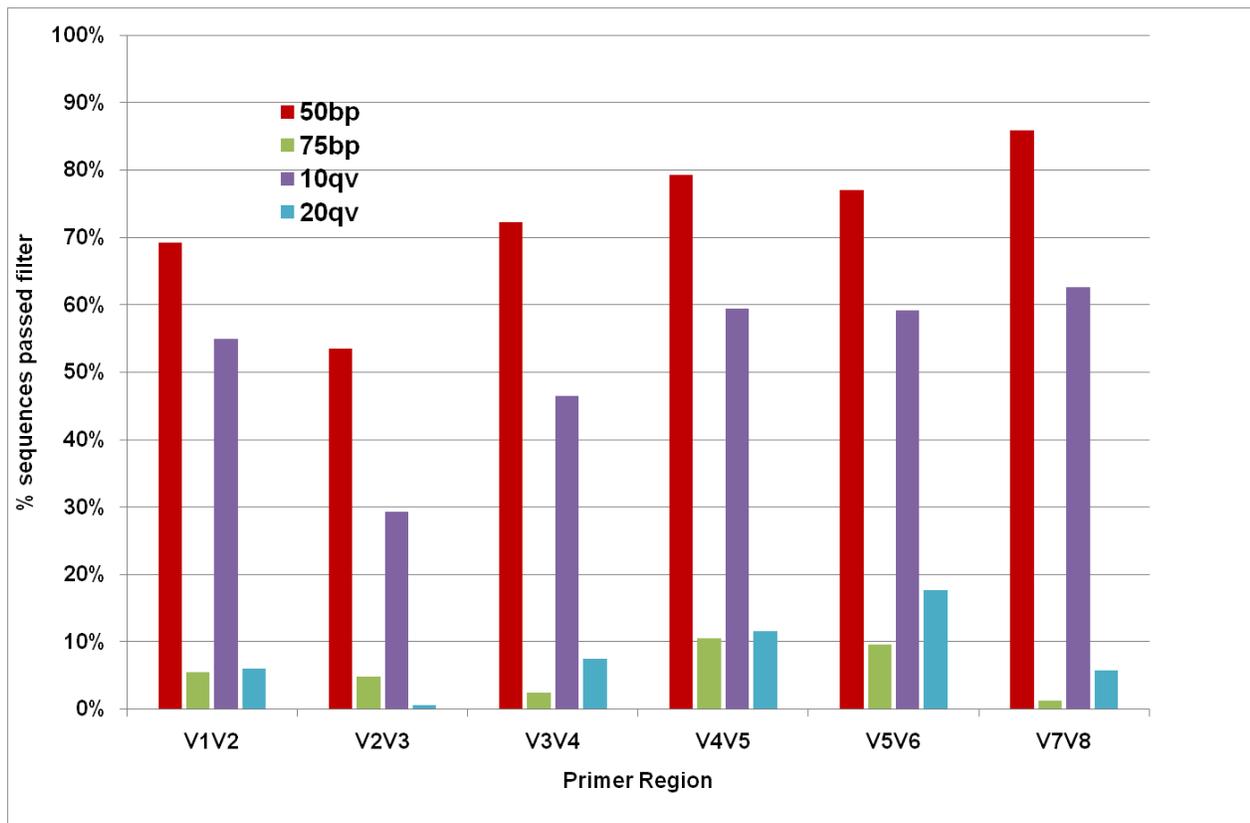


Figure S2. Phylum-level assignments of Illumina reads passing the tested filters for the six amplicons. The total for each bar indicates the percent of reads assignable using the RDP-Classifer with a bootstrap cut-off of 50%.

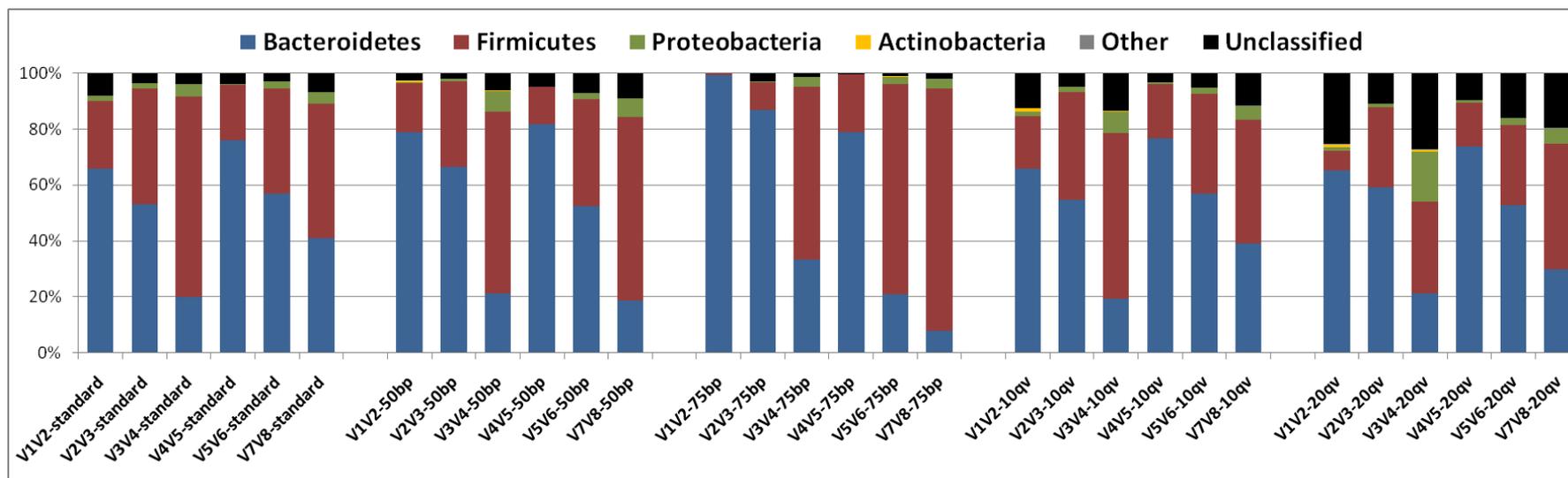


Figure S3. Relative phylum and genus abundances for three technical and one biological replicates.

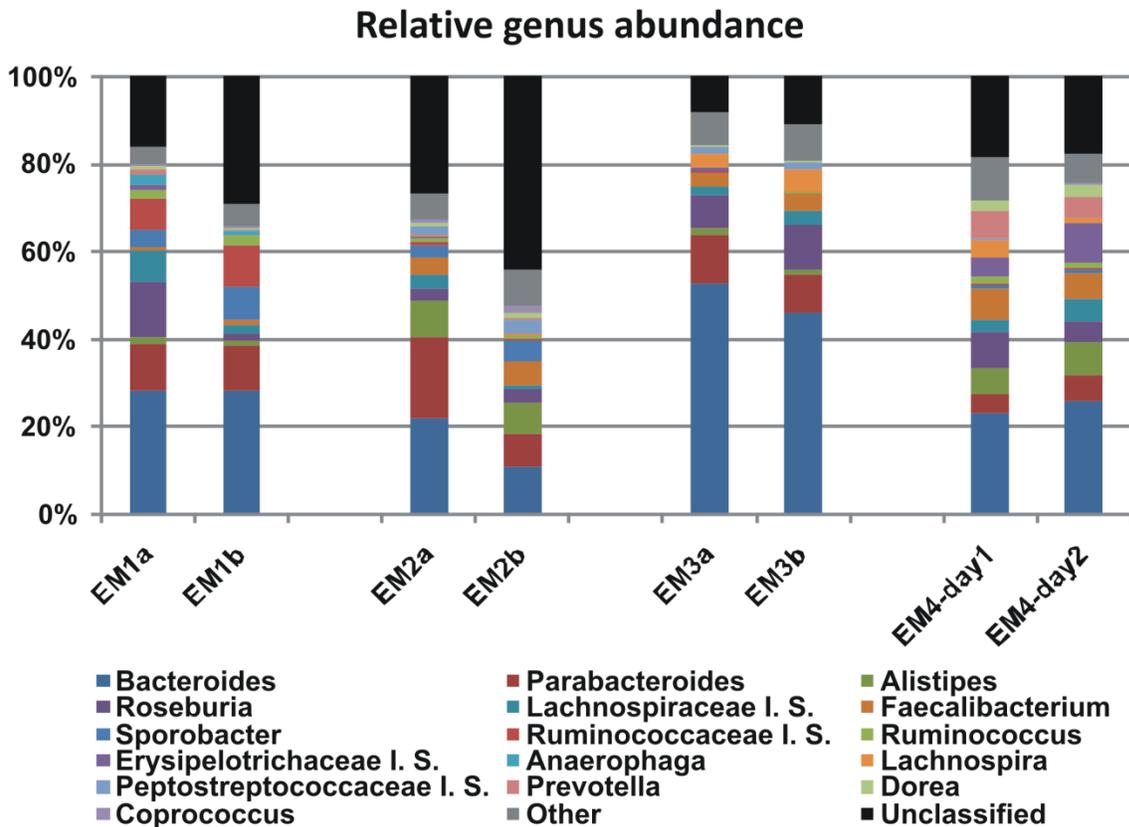
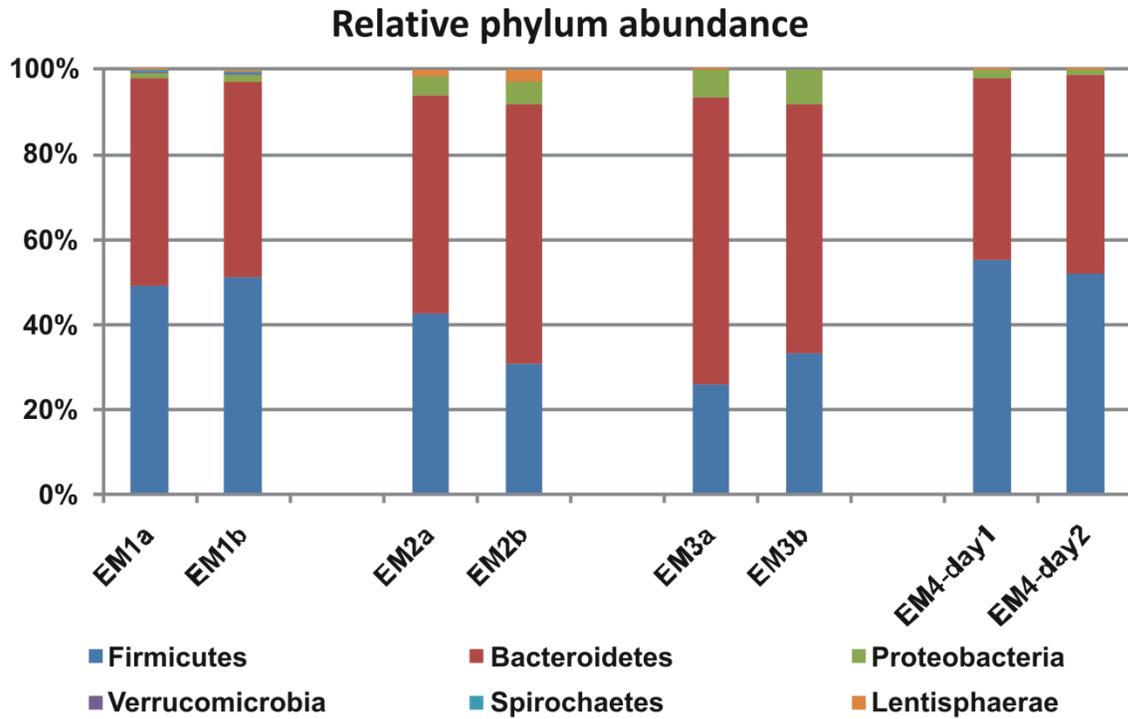
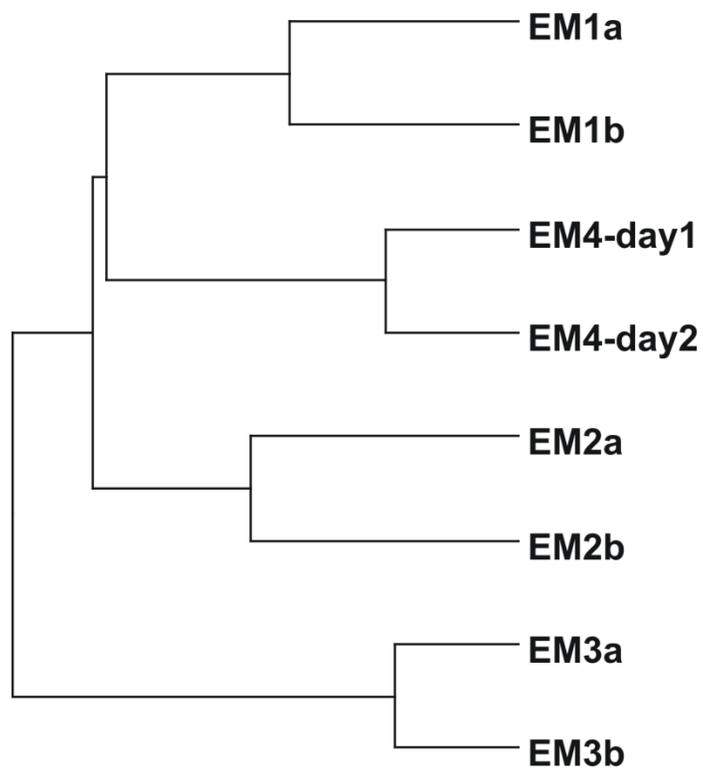


Figure S4. Hierarchical UniFrac clustering of three technical and one biological replicates using weighted (A) and un-weighted (B) UniFrac distances.

H0.01 **A**



B

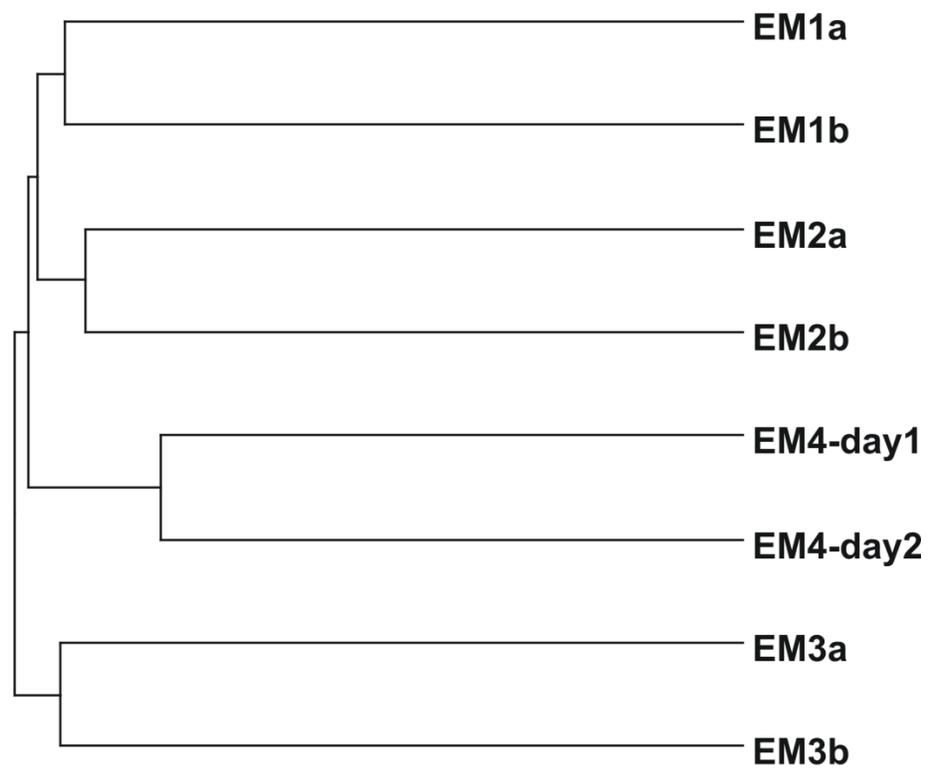


Figure S5. Proportion of full-length 16S rRNA and tandem regions from simulated Titanium and Illumina reads, accurately classified at four taxonomic levels. The reference set contains 60,000 high-quality full-length 16S rRNA gene sequences, not exclusively from the gastrointestinal tract. Sequencing errors were also introduced using error rates above (dashed lines: KIT-v4; dotted lines: KIT-v3).

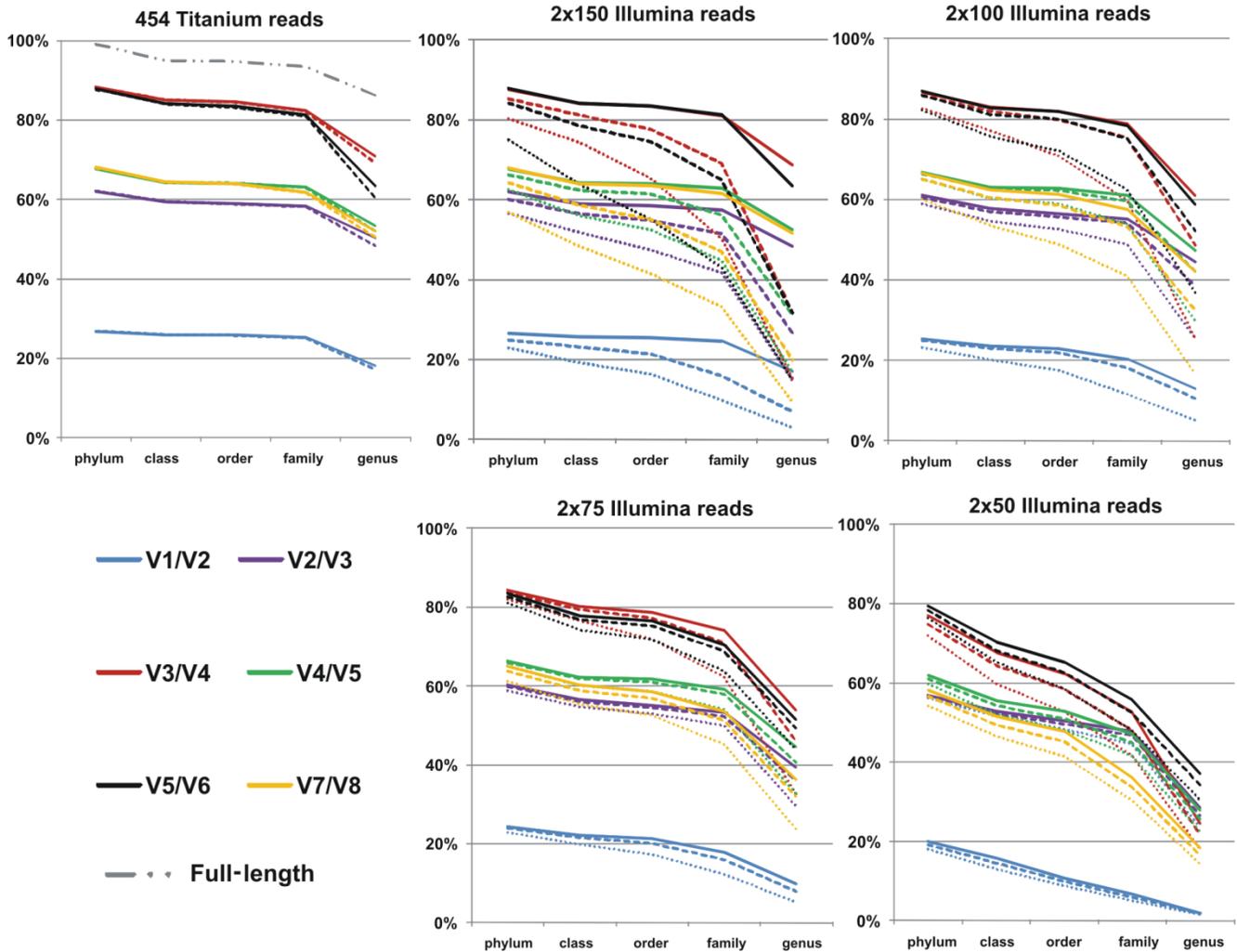


Figure S6. Percentage accurately classified single variable 16S rRNA gene regions.

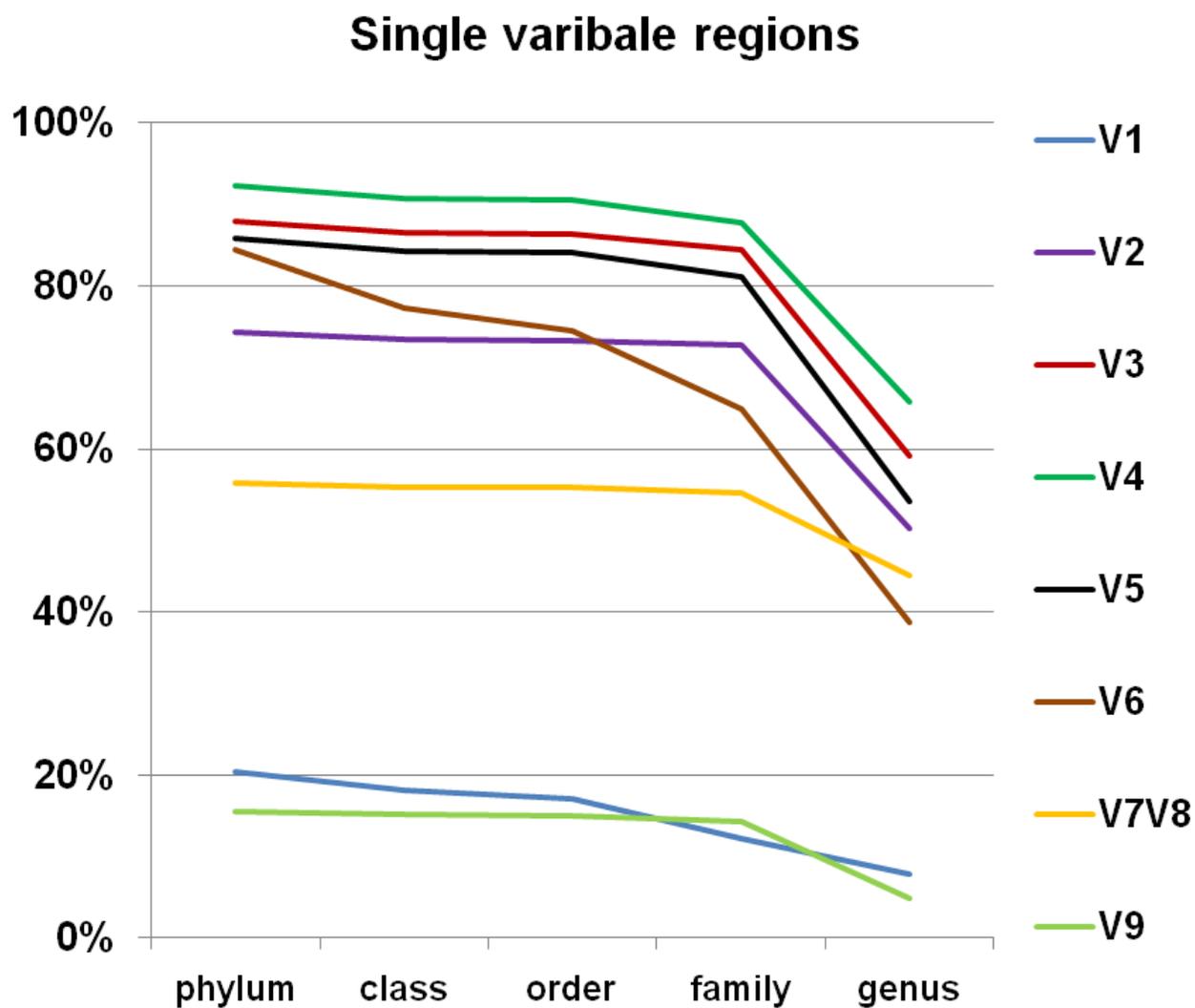


Figure S7. To further investigate reasons behind V3V4 and V7V8 deviations we compared family classifications between these amplicons and HITChip hybridisations of full-length 16S rRNA our previous study.

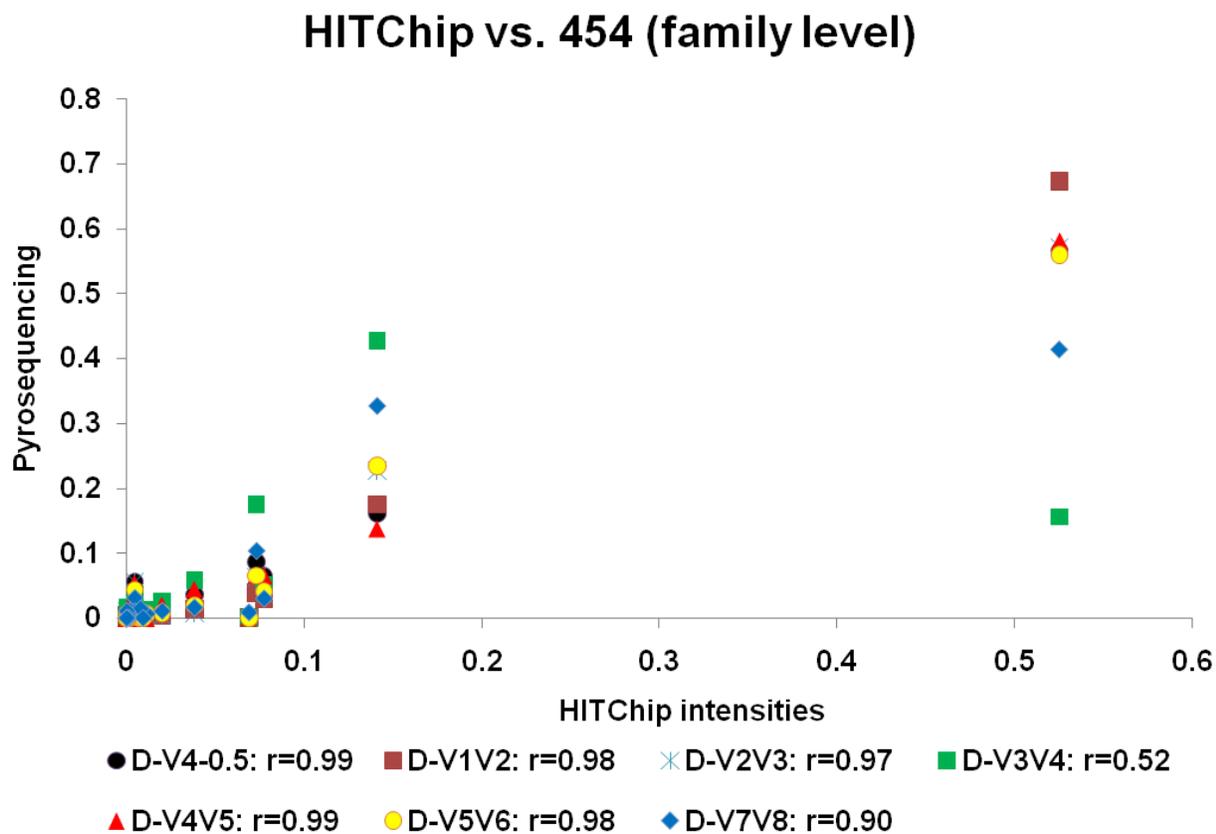


Figure S8. Correlations between Titanium reads assigned by the RDP-Classifer and MEGAN at phylum and genus levels.

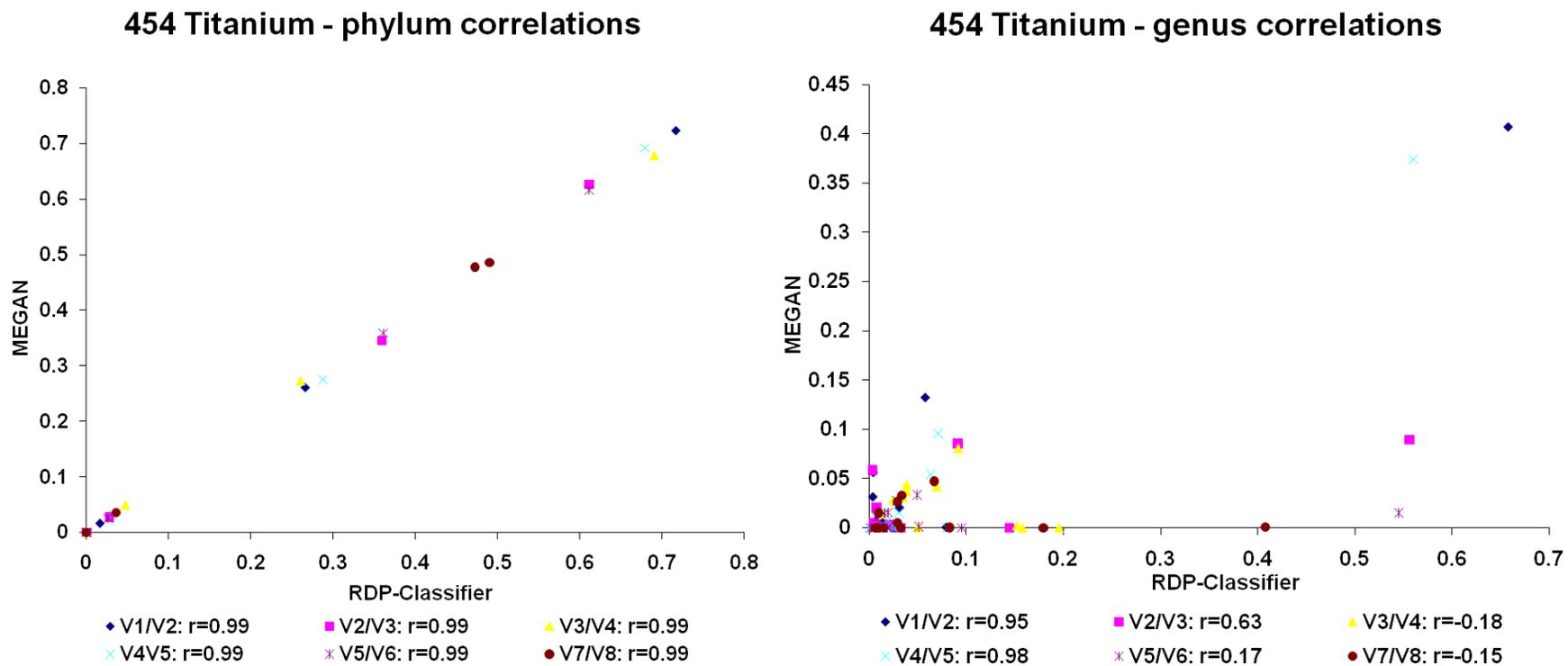


Figure S9. Correlations between Illumina reads assigned by the RDP-Classifer and MEGAN at phylum and genus levels.

