Published online 29 September 2010

Nucleic Acids Research, 2010, Vol. 38, No. 22 ***e200***
doi:10.1093/nar/gkq873

Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions

Marcus J. Claesson[1,2,*], Qiong Wang[3], Orla O'Sullivan[4], Rachel Greene-Diniz[1], James R. Cole[3], R. Paul Ross[2,4] and Paul W. O'Toole[1,2]

[1]Department of Microbiology, University College Cork, Cork, Ireland, [2]Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland, [3]Center for Microbial Ecology and Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA and [4]Teagasc, Moorepark Food Research Centre, Moorepark, Fermoy, Cork, Ireland

Received March 25, 2010; Revised September 9, 2010; Accepted September 16, 2010

ABSTRACT

High-throughput molecular technologies can profile microbial communities at high resolution even in complex environments like the intestinal microbiota. Recent improvements in next-generation sequencing technologies allow for even finer resolution. We compared phylogenetic profiling of both longer (454 Titanium) sequence reads with shorter, but more numerous, paired-end reads (Illumina). For both approaches, we targeted six tandem combinations of 16S rRNA gene variable regions, in microbial DNA extracted from a human faecal sample, in order to investigate their limitations and potentials. *In silico* evaluations predicted that the V3/V4 and V4/V5 regions would provide the highest classification accuracies for both technologies. However, experimental sequencing of the V3/V4 region revealed significant amplification bias compared to the other regions, emphasising the necessity for experimental validation of primer pairs. The latest developments of 454 and Illumina technologies offered higher resolution compared to their previous versions, and showed relative consistency with each other. However, the majority of the Illumina reads could not be classified down to genus level due to their shorter length and higher error rates beyond 60 nt. Nonetheless, with improved quality and longer reads, the far greater coverage of Illumina promises unparalleled insights into highly diverse and complex environments such as the human gut.

INTRODUCTION

Complex microbial communities, like the human gastrointestinal tract (GIT) and other bacterium-dense environments, are currently receiving increasing interest, due in large part to technological advances in culture-independent methods in recent years. Compared to capillary sequencing and non-sequence-based molecular methods, high-throughput sequencing provides unparalleled insight into community structures. Typically carried out by pyrosequencing on a 454 Genome Sequencer FLX machine (1), amplicons (sequence reads) of a single variable 16S rRNA gene region are quantified and subsequently assigned to microbial phylogenies (and thence to taxonomies). The nine different variable 16S rRNA gene regions are flanked by conserved stretches in most bacteria (2), and they can be used as targets for PCR primers with near-universal bacterial specificity (3,4). Although less discriminatory than the full-length 16S rRNA gene, massively parallel sequencing of the shorter reads offer either much higher coverage per sample (5) or many more samples per instrument run by means of innovative barcoding techniques (6,7). The trade-off with the longer, but fewer, reads generated by traditional capillary sequencing means a lower proportion of amplicons that can be classified at genus or species levels. In contrast, the resolution of the community composition with amplicon pyrosequencing is potentially several orders of magnitude

*To whom correspondence should be addressed. Tel: +353214901306; Email: mclaesson@bioinfo.ucc.ie

larger than clone library sequencing, and can be achieved at a significantly lower cost.

Different variable regions have been targeted in different studies. Generally, this selection has not been dependent on the sampled environment, but rather on published or unpublished recommendations and/or experimental familiarity with a certain region in the author's laboratory. A few comparative studies have focused on assessing region suitability: after using different methodological approaches Sundquist *et al.* favored the V1/V2/V4 regions (8); Liu *et al.* the V2/V3/V4 regions (9); Wang *et al.* the V2/V4 regions (10) and Chakravorty *et al.* the V2/V3 regions (11). Recently, we compared high and low pyrosequencing coverage of the V4 and V6 regions and concluded that the RDP-Classifier consistently assigned more V4 than V6 reads from the human GIT down to genus-level (5). Furthermore, a lower coverage (40 000 reads per sample) was sufficient to capture the majority of the bacterial diversity that was identified by five times greater sequencing depth. Pyrosequencing of the V4 region also yielded compositional profiles that were consistent with HITChip analysis (12), whereby we hybridised full-length 16S rRNA genes from two samples onto a phylogenetic array containing probes of concatenated V1 and V6 sequences.

As compositional studies like these depend on amplicon generation, they are subject to PCR bias of varying degrees (13,14). Although it was recently shown that amplicon length somewhat affected phylotype richness when comparing pyrosequencing reads of the V1–V2 and V8 regions of microbiota from the termite hindgut, the choice of region had a much larger impact on diversity values such as community richness and evenness (15). Moreover, parallel work from the same group suggested that far too relaxed quality filtering of raw pyrosequencing reads had been applied in many previous studies, thereby inflating previously reported diversity estimates measured at predefined phylotype similarity levels (16), as was also alluded to in an earlier study (17). However, pyrosequencing errors seem to have a lesser impact on both phylogenetic assignment rates (18) and methods comparing diversity across different communities (19). In a separate study, Wang and colleagues highlighted the lack of coverage and scope estimates for known 16S rRNA primers, and in response generated a comprehensive list of tested primers, along with recommendations of a few with particularly high coverage and universal properties (20).

Another massively parallel sequencing technology, that was first described the same year as 454 Pyrosequencing, is the Illumina technology [then Solexa Ltd. (21)]. Since then, the Illumina Genome Analyzer instruments have routinely been producing more than ten times the number of reads per run as the 454 GS FLX machines, albeit of much shorter lengths (typically between 36 and 76 bp). Lazarevic and colleagues sequenced over 1.3 million single reads of the latter size from the 16S V5 region, in order to explore the human oral microbiota (22). The higher coverage allowed the identification of low-abundance genera not detected in earlier studies of oral microbial flora. However, the limitations to the application of the Illumina technology for compositional studies were noted in that study, and future enhancements were predicted to increase its suitability for environments of even higher complexity. Recently, Illumina sequencing has also been applied to other single variable 16S rRNA regions, such as the V4 region in various environmental communities (23), or the V6 region in vaginal microbiota (24).

In this study, we took advantage of recent performance improvements in both the Illumina (paired-end 101 bp reads) and Pyrosequencing (>400 bp reads) technologies, and applied these on the human gut microbiota. We targeted the highly diverse microbial community within a single human faecal sample by separately sequencing both entire amplicons from six tandem variable 16S regions, and flanking ends thereof. In addition to evaluating the effects of biases imposed by targeting different variable regions and commonly used primers, we discuss the parameters for what may become the future methods of choice for microbial community composition analysis.

MATERIALS AND METHODS

High-throughput sequence assignment simulation

To explore the potential of microbial community composition analysis using Illumina and 454 Titanium sequencing, as well as how classification accuracy varies with reads of different length and quality, we compiled a high-quality reference set. This was based on the SILVA SSURef database release 100 (25), comprising 409 907 near full-length 16S rRNA sequences. To increase its quality and make it representative for bacterial communities within the GIT, the following filtering criteria were applied: (i) only bacterial sequences with no known anomalies, e.g. chimeras (pintail-score = 100); (ii) sequence length at least 1300 bp; (iii) existing RDP-classification not containing unclassified bacteria; (iv) sequence quality at least 90 (out of 100) and (v) isolated from samples of human or animal GIT or faeces. This filtering process resulted in a high-quality reference set of 27 013 full-length 16S rRNA sequences, whereof 98% originated from uncultured bacteria.

By using annealing locations for primers listed in Table 1, sequences for the six variable tandem regions were extracted *in silico* from the full-length sequence reference set, mimicking data from 454 Titanium reads (Figure 1). To also simulate paired-end and variously sized Illumina reads from the same 16S regions, 150/100/75/50 bp fragments were in turn extracted from both ends of these Titanium-length reads, filling the interior regions with 20 N residues. The RDP Probe Match program was used for calculating coverage among 16S rRNA sequences in the RDP database. In addition to simulating reads with perfect quality, we introduced stochastic errors along the read lengths according to error rates provided by the sequencing vendors for 454 Titanium (Figure 2a), and Illumina (average of data from forward and reverse reads) using the two most recent sequencing kits (Figure 2b). The simulated reads were then taxonomically

Table 1. Coverage of primers included in this study, both separately and in combination

Primer	Sequence	RDP Probe Match coverage (%)	Simulation coverage (%)	References
V1-forward	5′-AGAGTTTGATCCTGGCTCAG	64	42	8F/19; (38)
V2-reverse	5′-CTGCTGCCTYCCGTA	94	96	BSR357/15; (39)
V1/V2 combined		**64**	**40**	
V2-forward	5′-AGYGGCGNACGGGTGAGTAA	72	77	F101/19; (8)
V3-reverse	5′-ATTACCGCGGCTGCTGG	86	93	R534/17; (35)
V2/V3 combined		**60**	**72**	
V3-forward	5′-ACTCCTACGGRAGGCAGCAG	93	96	F338/19; (35)
V4-reverse	5′-TACNVGGGTATCTAATCC	90	96	R802/18; RDP website (http://pyro.cme.msu.edu/pyro/help.jsp)
V3/V4 combined		**86**	**91**	
V4-forward	5′-AYTGGGYDTAAAGNG	97	98	F563/16; RDP website (http://pyro.cme.msu.edu/pyro/help.jsp)
V5-reverse	5′-CCGTCAATTYYTTTRAGTTT	83	90	BSR926/20
V4/V5 combined		**81**	**88**	
V5-forward	5′-RGGATTAGATACCC	83	96	BSF784/15
V6-reverse	5′-CGACRRCCATGCANCACCT	94	97	R1064/18; (40)
V5/V6 combined		**87**	**93**	
V7-forward	5′-GYAACGAGCGCAACCC	88	89	BSF1099/16
V8-reverse	5′-GACGGGCGGTGWGTRC	87	61	BSR1407/16
V7/V8 combined		**84**	**56**	

Primer references prefixed with a BS notation were obtained from the European Ribosomal RNA Database.

assigned using the RDP-classifier (10) with a bootstrap cut-off of 50%. We had previously found this cut-off value to achieve the optimal balance of between achieved accuracy and retained number of reads (5). As the RDP-classifier ignores any 8-mer words with Ns, the interior regions have no impact on classification results. Classifications of both the simulated reads and of their originating full-length 16S rRNA sequences were imported into a MySQL database. This allowed fast and precise comparisons with the reference set, resulting in measurements of classification accuracy for each set of simulated reads.

Sample processing and sequencing

A faecal sample was collected from an 87-year-old female [subject D in our previous study (5)], who was a member of a larger cohort of elderly subjects recruited for the ELDERMET project (http://eldermet.ucc.ie). The Clinical Research Ethics Committee of the Cork Teaching Hospitals (CREC) granted full approval to the ELDERMET project on the 19th February 2008 [Ref: ECM 3 (a) 01/04/08]. Formal written consent was obtained, on the basis of an Information Sheet/Safety Statement, following an ethics protocol that was approved by CREC, in compliance with pertaining local, national and European ethics legislation and guidelines to best practice. The subject was taking an unknown antibiotic at the time of sampling. The sample was processed from fresh stool the same day as collection and DNA was extracted according to standard protocol (Qiagen, West Sussex, UK). Six amplicon libraries were created of variable 16S rRNA tandem regions using primers in Table 1. Standard PCR reaction conditions were employed for reactions with Taq polymerase: 2 mM $MgCl_2$, 200 nM each primer and 200 μM dNTPs. The PCR conditions were 94°C for 50 s (initialization and denaturing) followed by 40°C for 30 s (annealing), 72°C for 60 s in 35 cycles (extension), and a final elongation step at 72°C for 5 min. Two negative control reactions containing all components, but water instead of template, were performed alongside all test reactions, and were routinely free of PCR product, demonstrating lack of contamination with post-PCR product. The optimal annealing temperature for the primers, which included either the 454 adapters or the standard paired-end Illumina adaptors, was empirically determined by gradient PCR using control reactions with initially purified bacterial genomic DNA, and validated on faecal microbial community DNA (data not shown). The usage of region-specific 16S rRNA primers made additional barcodes redundant. All six amplicons were pooled and subsequently sequenced on a 454 Genome Sequencer FLX Titanium one-quarter picoliter plate (Cogenics, Essex, UK) according to 454 protocols. In addition, the same pool of samples was sequenced on one Illumina GA-IIx lane (Fasteris, Geneva, Switzerland) for 101 cycles from both ends of paired-end library preparations, using sequencing kit version 3.0 followed by base-calling using the GAPipeline version 1.4.0.

Sequence analysis

Raw pyrosequencing reads were quality trimmed according to published recommendations (26) using a locally installed version of the RDP Pyrosequencing Pipeline (27): sequences with inexact matches to both primer sequences, having poor quality, one or more ambiguous bases or read-lengths at least 20 bp shorter than the electropherogram peaks for each set of amplicon, were filtered off. Chimera sequences were detected with ChimeraSlayer (28).With the exception of the last criterion, the same criteria were applied to Illumina reads. Prior to this, purity filtering with 'chastity' values >0.6 and a maximum of one failed base call in the first 24 bases was applied to the raw reads. Additional filtering criteria were

also applied and evaluated (Supplementary data). Once filtered, the reverse Illumina read for each amplicon was reverse-complemented and merged with the corresponding forward read, inserting 20 Ns in between. Both forward and reverse 16S rRNA primers were removed from all pyrosequencing and Illumina reads. The primer sequences carry per definition very little phylogenetic information, so their removal did not have an adverse affect on taxonomic classifications. This was also supported by tests with and without primers on the simulated reference set (data not shown).

The Naïve Bayesian Classifier (RDP-Classifier) was used for assigning reads into the new Bergey's bacterial taxonomy (29) with a bootstrap cut-off of 50%. Trimmed sequences along with their classifications were imported into a MySQL database for efficient storage and advanced querying. To explore an alternative assignment method, a hierarchical tree summarising read assignments into the NCBI taxonomy was constructed using MEGAN (30) on BLAST searches against a previously published 16S rRNA—specific database (31) (with a bit-score threshold of 86, allowing ten hits per read). Pyrosequencing reads were aligned using Infernal (32) and associated covariance models obtained from the Ribosomal Database Project Group. Phylotype clusters of 97% similarity were obtained by applying the furthest neighbour approach using the Complete Linkage Clustering application of the RDP pyrosequencing pipeline. From these, rarefaction curves, Shannon diversities and Chao1 richness estimations were calculated using RDP software. Good's coverage was calculated as $G = 1-n/N$, where n is the number of singleton phylotypes and N is the total number of sequences in the sample.

RESULTS

In silico predictions show that classification accuracies are highly dependent on choice of region, sequencing technology and sequence quality

The complete sequences of six tandem variable regions, extracted from the 27 013 high-quality full-length 16S rRNA genes reference set, were used to simulate 454 Titanium reads, whereas the 150/100/75/50 bp of the flanking ends were used to simulate corresponding paired-end Illumina reads of varying lengths (Figure 1). The coverage of both single and paired primers, as measured by the RDP Probe Match and matches against the reference set (Table 1), were generally high except for the V1/V2, V2/V3 and V7/V8 regions, in large part due to poor coverage of the single V1-for, V2-for and V8-rev primers. Many full-length sequences used in the two reference sets have truncated ends, thereby lacking complete sequences covering the V1-for and V8-rev primer regions, which is a likely reason for poor coverage with these primers.

When ignoring sequencing errors, the 2×150 bp Illumina reads were almost as accurate as the longer Titanium reads, owing to their concatenated lengths approaching full Titanium read lengths (Figure 3). Not surprisingly, the genus-level accuracies for Illumina reads dropped as their read lengths decreased. Titanium reads were, however, still far from full-length 16S rRNA gene assignment accuracy. Regardless of sequencing technology and quality, the V3/V4 and V4/V5 regions were the most accurate. While *in silico*—induced errors, as modelled by error rates provided by the sequencing vendors (Figure 2), had little effect on the classification accuracies for pyrosequencing reads, they had a significantly negative impact for the longer Illumina reads which had increasingly deteriorating quality after 60 bp. Paradoxically, the longer Illumina reads were actually less accurate for genus-level assignment than the shorter ones, since error rates increase exponentially with read length. For example, although the accuracy for the V4/V5 regions increased with error-free read lengths, the 2×75 bp reads with induced KIT-v4 errors (dashed lines) were slightly more accurate at genus-level than the corresponding longer reads. With or without sequencing error, paired-end 50 bp Illumina reads are, however, not worth pursuing for these types of compositional studies.

Since this analysis was based on GIT-related 16S rRNA genes, we also wanted to investigate whether we would obtain similar results if this criterion was removed, i.e. by not restricting reference sequences to GIT environments. Consequently, the reference set was increased to 60 000 high-quality full-length sequences on which we repeated the simulations. As displayed in Supplementary Figure S5, accuracy values for all regions and read lengths are lower at all taxonomic levels compared to the GIT-restricted sequences. Interestingly, the V4/V5 region in particular showed inferior performance and was here marginally better than the V7/V8 region. The RDP-classifier is trained on well-characterized 16S rRNA genes sequenced from compositional studies of diverse microbial environments. As such, GIT-related microbiota are over-represented relative to non-GIT environments (e.g. 45%

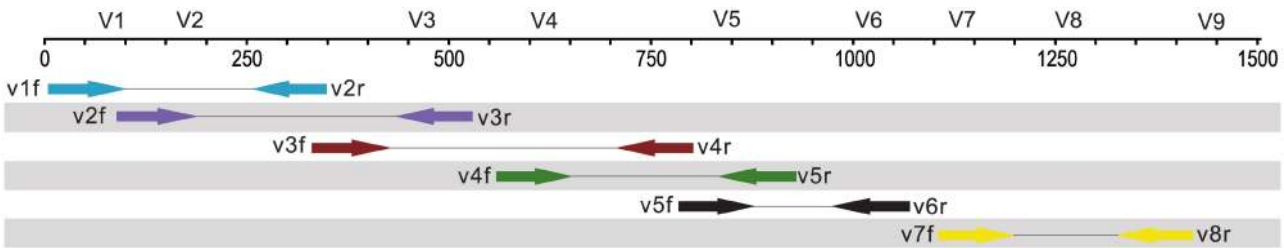

Figure 1. Positions of primer sequences and tandem regions used in this work for 454 titanium and Illumina, mapped along 16S rRNA gene (co-ordinates based on the *Escherichia coli* 16S rRNA gene sequence). The arrows (~100 bases) show approximately Illumina sequence read length.

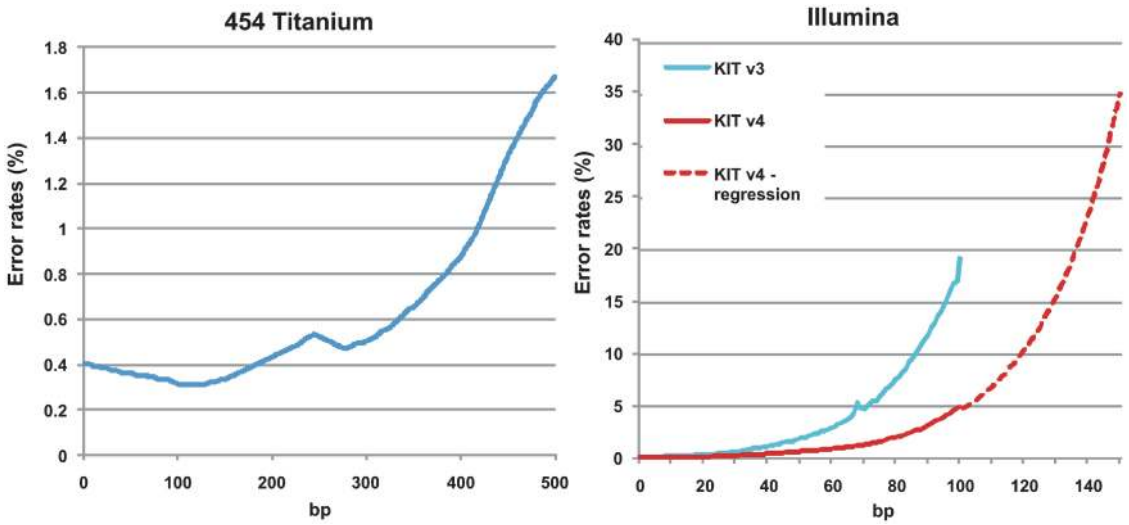

Figure 2. Error rates as function of read lengths (provided by Roche and Fasteris). Error rates beyond 100 bp for Kit v4 were obtained through extrapolation [$f(x) = 0.0763e^{0.0408x}$].

of the 60 000 reference high-quality full-length sequences originate from the GIT) which could thus explain the different patterns in classification accuracy.

As a reference, we also calculated similar accuracies for single variable regions (Supplementary Figure S6). Reassuringly, the accuracy of single V3/V4/V5 regions were consistent with, and slightly lower than, corresponding tandem regions of Titanium length derived from simulations of the GIT-related reference set. The V1 and V9 regions showed the poorest results, followed by the V7/V8 region.

Diversity metrics are highly dependent on region choice and sequencing depth

The same regions modelled above were sequenced on both the 454 Titanium and Illumina platforms, using primers from Table 1. Our filtering approach decreased the numbers of accepted reads significantly, notably more for the Illumina reads. As quality deteriorates dramatically with increasing read length for Illumina (Figure 2b), we investigated the effect of additional quality filtering criteria (Supplementary Data). We concluded that the standard criteria provided the best balance of good classification efficiency, high retention of reads, even composition between regions and high similarity to composition as derived by 454 Titanium reads.

Sequencing and diversity characteristics (based on the 97% phylotype similarity level) are outlined in Table 2. The variations in amplicon mean lengths for 454 Titanium is a reflection of the differently sized tandem regions, while Illumina reads are consistently of the same length (2×101 bp). Length deviations for the latter technology are instead due to trimming of the variously sized primer sequences. The observed richness levels varied dramatically depending on sequenced region and adapted technology; for Titanium reads, the range was 349 to 1146 phylotypes, and from 97 670 to 173 857 phylotypes for Illumina reads. Interestingly, the richness values that were the highest/lowest for Titanium were not necessarily the highest/lowest for Illumina reads. This is probably because the shorter Illumina reads sometimes cover regions of variability different from the overall variability as sequenced by the longer Titanium reads for the same region.

Figure 4A shows similarly deviating rarefaction curves for the six different 454 Titanium and Illumina amplicons. Similar curves for the complete set of Illumina reads were omitted due to computational difficulties, and for their apparent lack of reliability. Instead we calculated curves from random sub-samplings of 229 048 reads, equal in size to the region with fewest reads. We included curves for random sub-samplings of 8277 reads (amplicon V2/V3) to examine the underestimating effect we had previously observed (5), and which was also pronounced here for all amplicons except V4/V5. The inflated richness levels for the Illumina sequenced amplicons, as well as the nearly linear rarefaction curves, seemed unrealistic at best and are presumably artefacts of the high error rates in combination with the vast number of reads for each amplicon. This is also supported by the fact that richness values derived from random sub-samplings of Illumina reads (equal in numbers to the corresponding 454 regions) were substantially higher than for the 454 reads at the same read number levels (Table 2 and inset Figure 4A). Likewise, Good's coverage values from Illumina reads are relatively small compared to the corresponding Titanium reads. This parameter is an estimator of the completeness of sampling (33) and should not be mistaken for sequencing coverage. High error rates produces many singleton phylotypes which results in lower Good's coverage values (see formula in Methods). Thus, for the same reasons that the rarefaction curves are nearly linear and the Chao1 richness estimations are extremely high, Good's coverage is relatively low; the

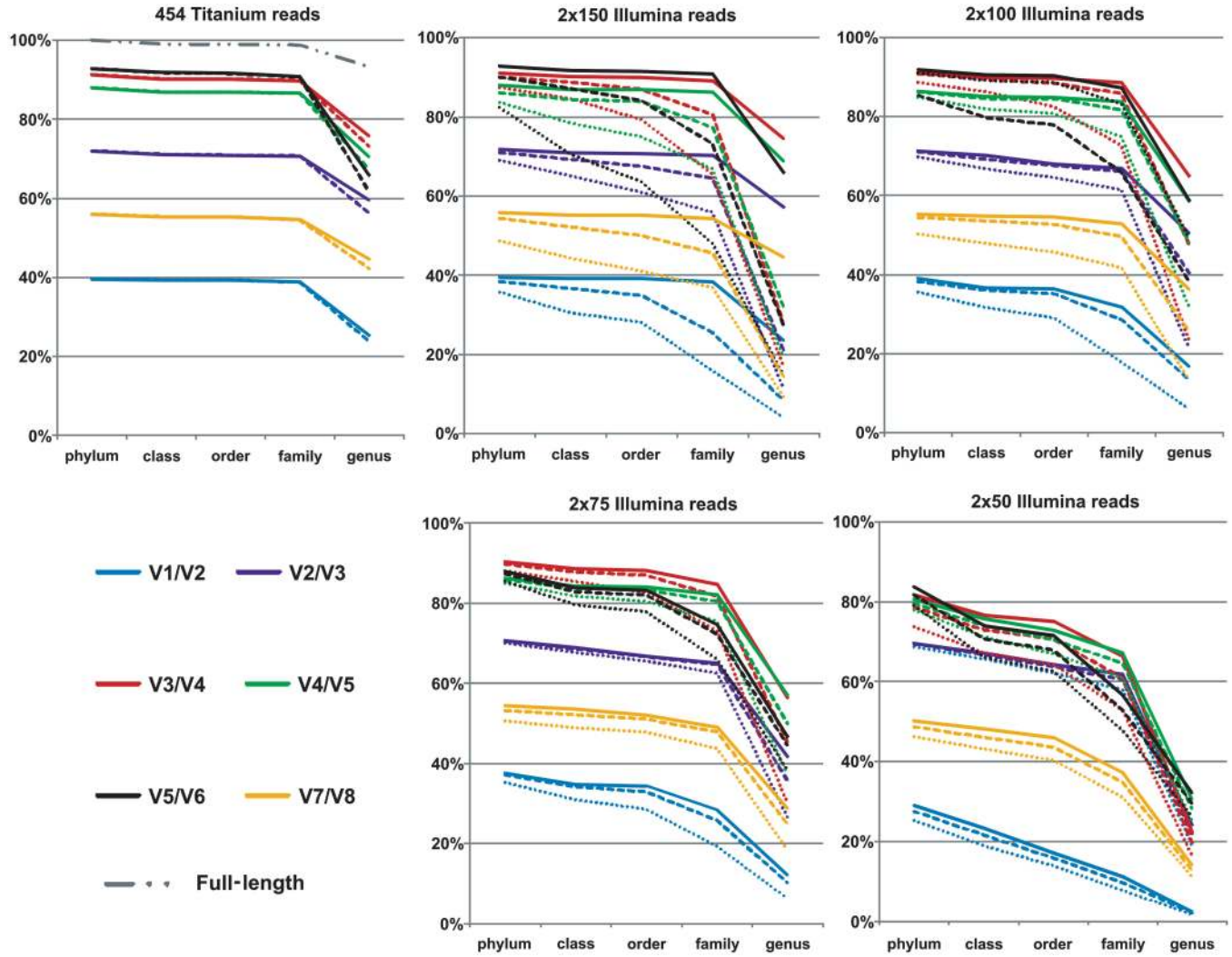

Figure 3. Proportion of full-length 16S rRNA and tandem regions from simulated Titanium and Illumina reads, accurately classified at five taxonomic levels. Sequencing errors were also introduced using error rates above (dashed lines: KIT-v4; dotted lines: KIT-v3).

V3/V4 and V7/V8 amplicon data produced the highest richness estimations and also the lowest Good's coverage values.

Classification efficiencies among sequencing technologies are significantly different and ultimately affect resolution

When comparing classification efficiencies (CE), defined as the proportion of all reads confidently classified to a certain taxonomic level (genus-level from here on), we observed acceptable values (>87%) for all Titanium reads (Figure 4B). We included reads from the single V4 region (5) as a comparison, and its CE was grouped in the middle of the variable tandem regions, below V4/V5 and, interestingly, above V3/V4. The Illumina technology had far worse CE for all regions, owing to a combination of shorter read lengths and poorer quality. Overall, the V4/V5 region showed best performance for both technologies, while the V3/V4 region was the worst.

It has been reported that in some studies that reverse reads on the Illumina instrument are inferior in quality to the forward reads (34). Here, however, we only noted a slight difference in the average quality values of 15.4 and 16.6 for forward and reverse reads, respectively. To investigate if this quality difference had an effect on the classification potential, we RDP-classified the forward and reverse reads separately, and found that there was a clear difference between forward and reverse reads from all six tandem regions in terms of representative sequences in the RDP reference database: Classification efficiencies at the genus level were between 12 and 36 percentage points higher for the forward reads than for the corresponding reverse reads. It is unclear whether these discrepancies are due to the slight quality difference between forward and reverse reads, or the fact that the partial variable regions covered by the reverse reads are all less discriminatory. However, the fact that the phenomenon was evident for all six tandem regions suggests that the former explanation is the more likely one.

In order to compare resolution levels provided by the two sequencing technologies, we quantified the number of unique genera that could be identified by the

Table 2. Amplicon sequencing characteristics for the six variable tandem regions

Amplicons	V1/V2		V2/V3		V3/V4		V4/V5		V5/V6		V7/V8	
Technology	454	Illumina	454	Illumina	454	Illumina	454	Illumina	454	Illumina	454	Illumina
Untrimmed reads	54 254	NA	26 132	NA	51 902	NA	42 438	NA	57 433	NA	62 040	NA
Trimmed reads	25 906	758 467	8277	556 866	21 492	229 048	12 223	481 594	26 399	887 427	35 531	1687 876
Mean Length ± SD	320 ± 6	167 ± 0.1	389 ± 10	165 ± 0.1	408 ± 18	164 ± 0.4	326 ± 3	167 ± 0.3	245 ± 8	169 ± 0.3	277 ± 6	170 ± 0.1
Phylotype richness	629 (338)	135 768 (17 169)	859	127 144 (5559)	696 (379)	164 566 (17 054)	349 (287)	111 205 (7180)	1146 (563)	97 670 (13 063)	980 (409)	173 857 (28 864)
Chao1 richness estimation	955 (560)	666 851 (107 596)	1889	446 174 (27 626)	1182 (645)	1023 682 (137 089)	611 (605)	436 606 (33 726)	2076 (1111)	290 606 (51 229)	1597 (815)	1179 242 (254 062)
Chao1-LCI95	863 (480)	655 501 (101 920)	1651	439 700 (25 419)	1055 (555)	1005 935 (128 972)	521 (483)	429 257 (31 385)	1887 (957)	286 245 (48 816)	1457 (684)	1158 578 (241 882)
Chao1-UCI95	1081 (685)	678 448 (113 653)	2198	452 782 (30 077)	1353 (779)	1041 805 (145 794)	749 (804)	444 124 (36 293)	2315 (1324)	295 067 (53 805)	1778 (1009)	1200 339 (266 940)
Shannon diversity	3.46 (3.40)	11.03 (9.11)	4.27	11.06 (8.15)	4.26 (4.15)	11.64 (9.48)	3.22 (3.19)	10.57 (8.17)	4.03 (3.91)	10.23 (8.39)	0.60 (3.95)	11.76 (10.03)
Phylotype evenness	0.538 (0.584)	0.933 (0.934)	0.631	0.941 (0.946)	0.651 (0.699)	0.969 (0.972)	0.549 (0.56)	0.910 (0.920)	0.573 (0.617)	0.891 (0.885)	(0.66)	0.975 (0.977)
Good's coverage	99.0% (98.3%)	50.9% (42.5%)	94.2%	56.9% (43.2%)	98.5% (97.9%)	37.6% (28.3%)	98.7% (98.3%)	62.1% (51.5%)	97.8% (96.5%)	68.7% (60.5%)	98.8% (97.5%)	32.9% (25.9%)

The number of untrimmed amplicon reads for separate Illumina amplicons is unknown as primer sequences were used for both separation and trimming. The values in parentheses in the 454 columns represent diversity metrics calculated on random sub-samplings of 8277 sequences, and the values in parentheses in the Illumina columns represent diversity metrics calculated on random sub-samplings equal in size to the corresponding 454 regions. Illumina values without parentheses were calculated from random sub-samplings of 229 048 reads equal in size to the region with fewest reads (the V3/V4 region). Diversity metrics are at the 97% similarity phylotype level.

RDP-classifier for the single V4 and tandem V4/V5 region (Figure 5). Even though the V4/V5 region were pyrosequenced with just over a quarter of the reads used for the single V4 region, 74% of all the V4 genera could be captured by the longer reads. Furthermore, the significantly higher Illumina coverage resulted in only a disproportionate increase in genera identified not detected by Pyrosequencing: These were *Sporobacterium*, *Paludibacter*, *Oribacterium*, *Campylobacter*, *Abiotrophia* and *Johnsonella*. This demonstrates that resolution is ultimately dependent on not only sequence coverage, but also classification efficiency, i.e. choice of region and sequence quality.

Relative taxa abundances are generally consistent across technologies, but show dramatic variation for two regions due to significant amplification bias

The two sequencing technologies revealed relatively similar profiles at phylum level, while they were more different at genus level (Figure 6). This is probably due to the much lower CE for Illumina, manifested by the significantly larger numbers of unclassified genera. It is not unlikely that genera which are classified to a higher extent with Titanium reads, such as *Lachnospiraceae Incertae Sedis*, are found within the larger cohort of unclassified Illumina reads.

Since the different sequencing targets and technologies used in this study were all applied on a single sample, we also wanted to investigate how phylum and genus profiles varied between replicates. Based on pyrosequencing of duplicate V4 amplicons libraries from four separate individuals (Supplementary Data), we found that even though taxonomic profiles between samples were not identical at phylum/genus level, all replicates still group together when compared to each other at the finest possible level of resolution (unique sequences). Thus, seemingly large variations in e.g. phylum distributions between samples may not necessarily reflect large differences in the overall microbiota, which should be taken into consideration when comparing the slighter variations in taxonomic profiles observed between the six sets of amplicons (Figure 6).

Intriguingly, the relative taxa abundances from the previously sequenced V4 region (5) were much more similar to V4/V5 than the V3/V4 region. The V3/V4 region also had by far the most deviating composition profile compared to the other regions, followed to some extent by V7/V8. This discrepancy was observed across the two technologies at both phylum and genus levels. Neither the RDP Probe Match (86% coverage) nor simulation (91% coverage) estimates (Table 1) implied any bias for the V3/V4 region. Similarly, Pearson correlations between genus classifications of full-length 16S rRNA sequences and the *in silico*—extracted variable regions from the reference set did not reveal any such bias either: V1/V2 ($r = 81\%$), V2/V3 ($r = 96\%$), V3/V4 ($r = 98\%$), V4/V5 ($r = 97\%$), V5/V6 ($r = 83\%$), V7/V8 ($r = 93\%$), and single regions V3 ($r = 85\%$) and V4 ($r = 87\%$). To further investigate reasons behind V3/V4 and V7/V8 deviations, we compared family classifications between

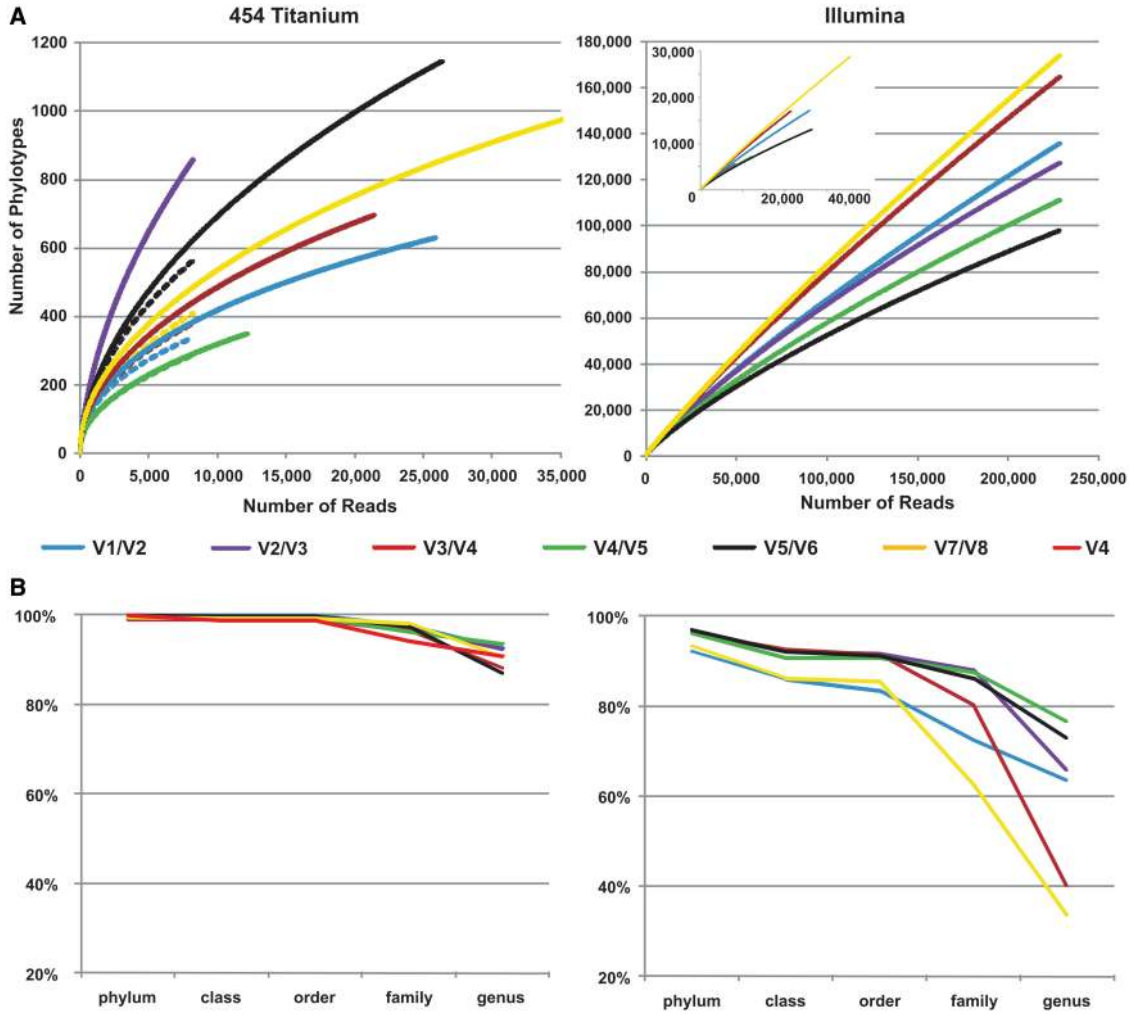

Figure 4. Rarefaction curves for Titanium and Illumina reads at the 97% similarity phylotype level. Dashed lines are for 8277 randomly sub-sampled Titanium reads, equal in size to the smallest Titanium amplicon dataset. Illumina rarefaction curves were calculated from random sub-samplings of 229 048 reads, equal in size to the region with fewest reads (the V3/V4 region). The inset shows rarefaction curves from randomly sub-sampled Illumina reads equal in numbers to the corresponding 454 regions. (**A**) Proportion of sequenced Titanium and Illumina reads that were classified at four taxonomic levels (**B**) Single V4 reads sequenced in our earlier study (5) were included for comparison.

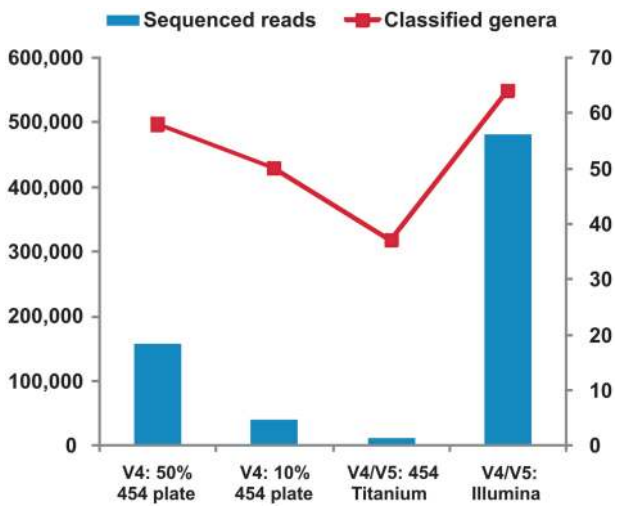

Figure 5. Resolution at genus level for Titanium and Illumina V4/V5 reads. Single V4 reads sequenced at two different depths in our earlier study (5) were included for comparison.

these amplicons and HITChip hybridisations of full-length 16S rRNA from our previous study (5). From Supplementary Figure S7 it is evident that both these regions, and especially V3/V4, have much poorer correlations with HITChip hybridisations. Following comparisons with the sequenced single V4 region we already know that the V4-rev primer used is not responsible for the said bias. To finally exclude the possibility that the V3-for primer is the sole error-causing source we compared aggregates of full-length 16S rRNA gene sequences with V3 reads, sequenced with capillary Sanger sequencing and 454 Pyrosequencing, respectively (35). The ratios of the two largest phyla *Bacteroidetes* and *Firmicutes* were 0.66 for the full-length sequences and 0.77 for the V3 reads, thus relatively close, and definitely not as disparate as for the V3/V4 and V4/V5 regions. Only 41 chimera sequences were detected among the V3/V4 Titanium reads, which again would not explain the observed difference. Altogether, these data are conclusive evidence that

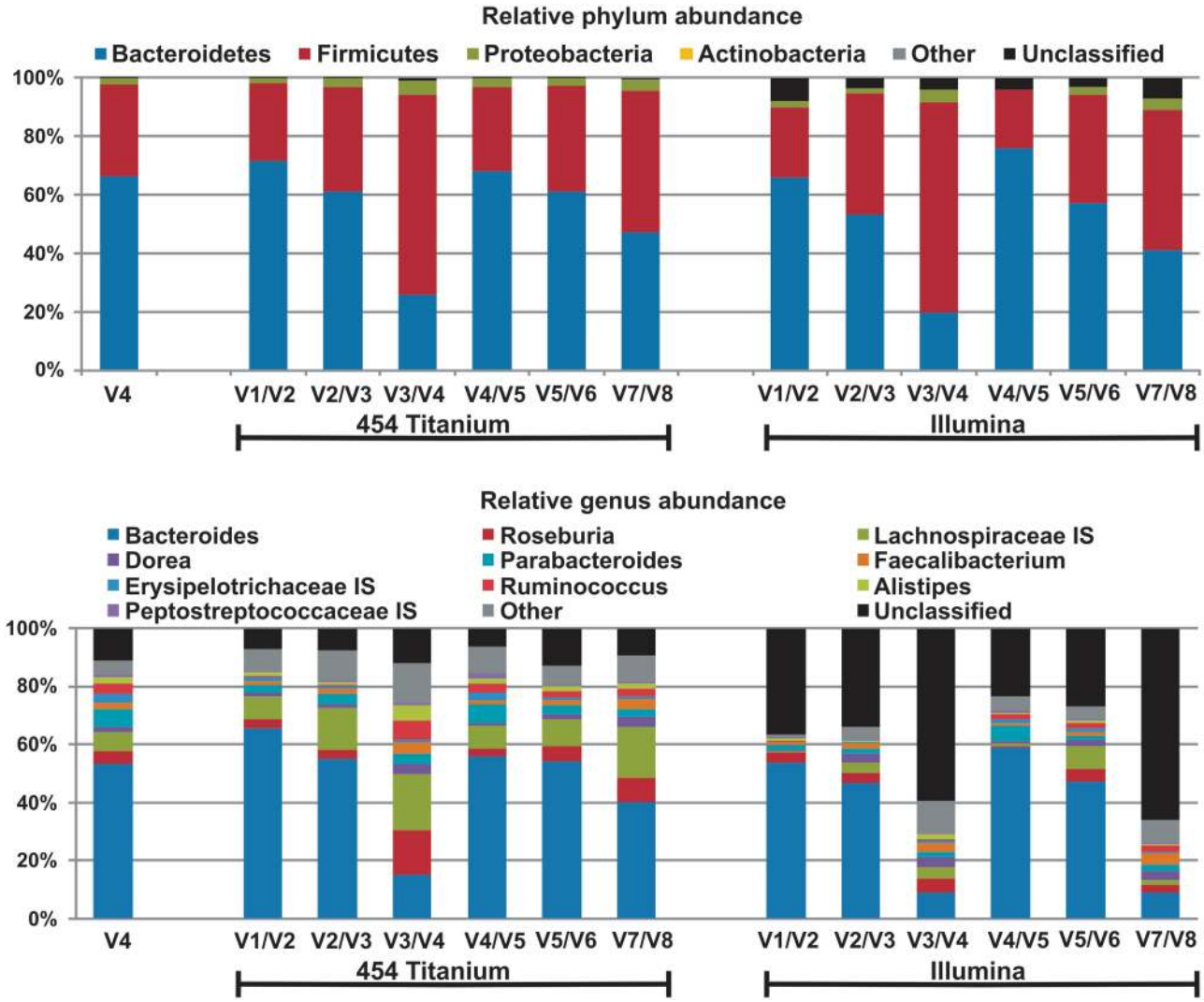

Figure 6. Relative phylum and genus abundances for sequence reads from both sequencing technologies. Single V4 reads sequenced using non-Titanium pyrosequencing in our earlier study (5) were included for comparison.

the V3/V4 deviations are due to bias associated with the experimental amplification process occurring when these particular V3-for and V4-rev primers are combined, rather than to uneven primer coverage.

MEGAN assignments are consistent with the RDP-classifier only for a minority of tandem regions

The RDP-classifier uses Bayesian probability theory for observing eight-character sub-sequences within each unknown query sequence, and has been trained on over 7000 bacterial full-length 16S rRNA genes. To investigate if a common alternative assignment approach would generate similar results, we applied the MEGAN tool (36) on BLAST searches of the trimmed Titanium reads against the RDP database. We deemed BLAST searches of the 4.6 million Illumina reads as being too computationally intense, and therefore performed the analysis on subsets of 40 000 sequences per region instead. Correlations between genus classifications of 10 different sub-samplings were all consistently high ($r > 0.99$), suggesting that any of these sub-sampled sets were representative for the complete set of Illumina reads. Figure 7 shows the comparison tree generated by MEGAN using all the Titanium reads and the subsets of Illumina reads, with relative taxonomic abundances for the six variable tandem regions at various taxonomic levels. Similarly to results from the RDP-classifier, composition profiles based upon the V3/V4 and V7/V8 regions indicated larger proportions of the *Firmicutes* phylum for both sequencing technologies. In contrast, there are several significant differences between the two assignment approaches at genus level; perhaps most strikingly, *Bacteroides* reads account for a large fraction of the community only for the V1/V2 and V3/V4 regions according to the MEGAN analysis of the Titanium reads.

To further investigate these discrepancies we generated correlation plots (Supplementary Figures S8 and S9) between phylum and genus classifications for the two approaches and sequencing technologies. For Titanium, only the V1/V2 and V4/V5 regions showed good

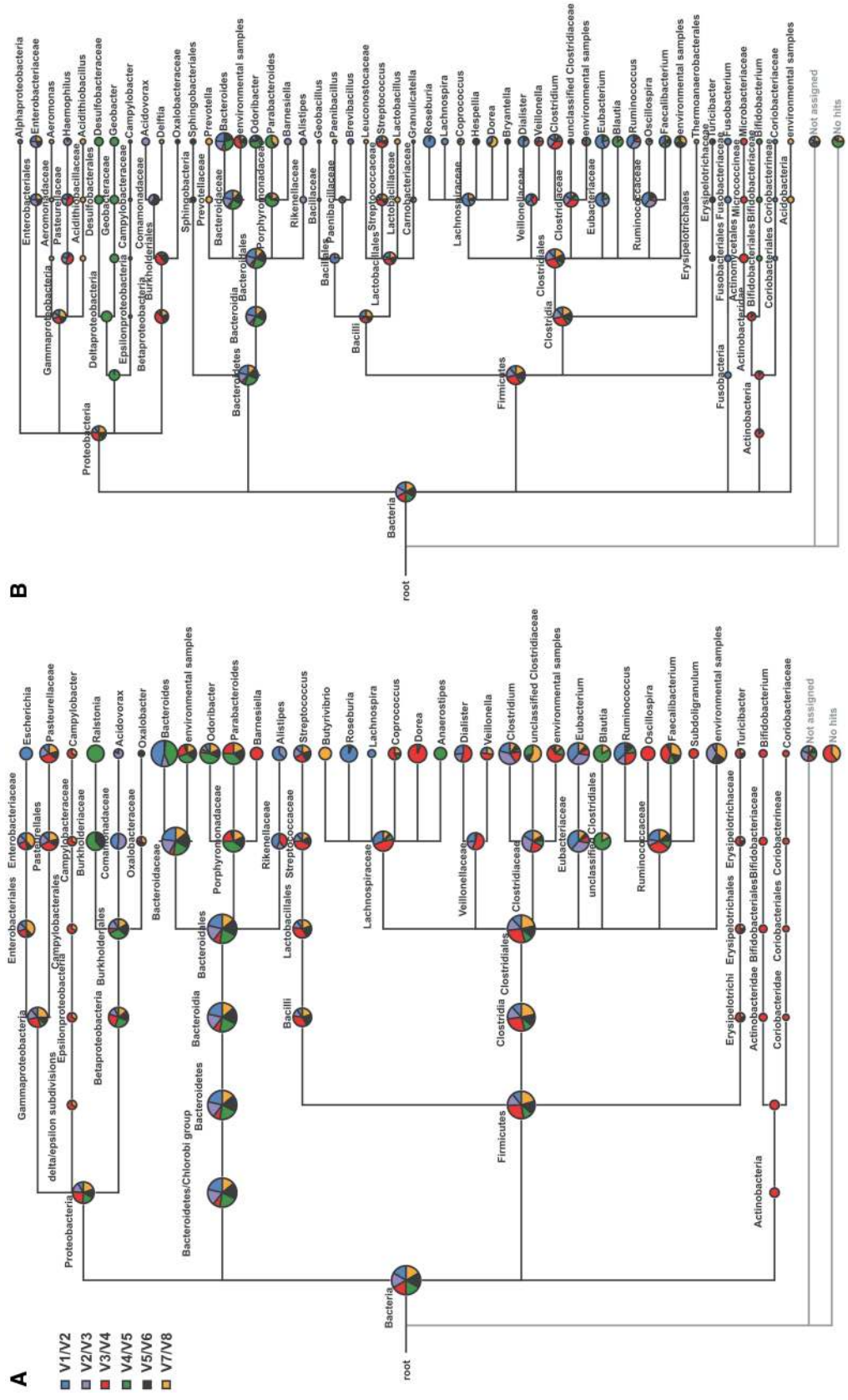

Figure 7. MEGAN comparison based on BLAST searches of all 454 Titanium reads (**A**) and a random subset of Illumina reads (**B**) against an rRNA-specific database.

correlations between the two classification methods, with Pearson correlations of 0.97 and 0.98, respectively. The reason behind the genus discrepancy was revealed from closer examininations of the MEGAN data for the *Bacteroides* assignments; in order to assign a read to the *Bacteroides* genus, all 10 first BLAST hits had to be against *Bacteroides* species. As it were in many cases, the ninth hit was against a group of bacteria labelled 'uncultured bacterium adhufec' (acronym for adult human faeces). These bacteria were, however, classified as belonging to the *Bacteroidales* family, and were, according to additional BLAST searches, unambiguous *Bacteroides* species (data not shown). Moreover, BLAST hits against the genera *Clostridium*, *Roseburia* and *Ruminococcus* are in many cases indistinguishable, which thus explain these genus deviations. In comparison, MEGAN analysis of the Illumina reads showed better consistency with the corresponding RDP-classifications, especially for the *Bacteroides* genus (Figure 7B and Supplementary Figure S6). The problematic ninth BLAST hit against the incorrectly labelled *Bacteroides* species was simply not an issue for the Illumina reads, since the reads had fewer hits with high scores. It is also important to note that the average classification efficiency for the RDP-classified Illumina reads was nearly twice that for the reads classified with MEGAN (59 versus 30%). To summarize, the deviating compositions of the V3/V4 and V7/V8 reads did not seem to be caused by poor performance of the RDP-classifier relative to the MEGAN approach.

DISCUSSION

The number of compositional studies of complex microbial communities that use high-throughput sequencing of partial 16S rRNA amplicons is increasing rapidly, encouraged by earlier successful studies and by the growing output-per-cost-ratio. Nonetheless, to obtain as accurate results as possible it is of paramount importance to minimize the amplification bias inherent in this approach, and to select variable 16S rRNA gene regions and sequencing primers with utmost care. Our main aim in this comparative study was not to investigate the primers with the highest performance expected, nor to test as many as possible. It was rather to investigate anomalous data generated with previously published primers, while at the same time evaluating their suitability, in new variable region combinations, in conjunction with recent sequencing technology improvements.

For sequencing by synthesis on the Illumina platform, standard paired-end linkers were ligated to the amplicons that been generated by universal 16S rRNA gene variable region primers. This does not significantly affect the yield of sequence data. Although it is theoretically possible that the ligation step might introduce a bias, such an effect has not been noted in the multiple genome re-sequencing projects completed on this platform (Fasteris, personal communication). Furthermore, analysis of the first base sequenced in any particular Illumina run did not identify bias towards a particular nucleotide (data not shown),which would be expected if there was a bias in the ligation, and the GC bias of the genome was maintained.

Based on simulation accuracies, classification efficiencies and consistency between two different classification approaches (RDP-Classifier and MEGAN based on BLAST searches), the V4/V5 region showed the best performance across the two sequencing technologies. Somewhat surprisingly however, we noted that sequenced reads of the V3/V4 region performed the worst; this was in spite of its high simulated accuracy (primer coverage and regional classification potential), and previous indications of good classification consistency for its constituent V3 and V4 parts (5,35). Hence, the bias was not associated with the selected individual primers or with the choice of sequencing method, but rather with amplification artefacts arising from the combination of these two specific V3-forward and V4-reverse primers. This emphasises that we should not blindly trust *in silico* predictions or primers, nor known results from separate components of the variable region in question. In contrast, support from actual amplification experiments using the proposed primer combination is absolutely necessary.

Moreover, even with longer variable regions, further developed sequencing technologies and higher coverage, it was evident that the microbial diversities measured from the same sample differed significantly depending on choice of variable region(s). We could therefore confirm the highly region-specific behaviour across datasets observed by other groups (16,17,37), and thereby re-iterate the weakness of comparing diversities between communities based on different ribosomal gene regions. Comparisons of additionally sequenced V4 amplicons also highlighted that although microbiota compositions may not be identical at phylum and genus level, their overall composition revealed at finer resolution could still have better discriminatory effect.

The extremely inflated diversity metrics, as derived from the Illumina reads, could in large part be explained by the high error rates above 60 bp. The exponentially deteriorating quality after this point was also the source of poor accuracy and classification efficiency for the shorter Illumina reads. It is possible that a more suitable alternative to these paired-end reads, which flank the variable tandem regions, could be shorter-insert fragments where the poor quality read ends partly overlap, resulting in improved consensus quality in the critical sequence region. At present, neither taxonomic classifications nor community diversity as derived from Illumina reads are reliable enough, and the coverage improvement over pyrosequencing does not result in an equivalently increased insight into the rare community members. Subsequent analysis of beta diversity (between subjects or along time series) would also produce unreliable results, due to this limitation. Notwithstanding, the technology has enormous potential, and when quality improves further, the Illumina technology may reveal unprecedented diversity from even the most complex microbial environments on earth.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Nessa Gallwey, Karen O'Donovan and Ann O'Neill for technical and clinical help, and to Siobhan Cusack for project management. This study is an output of the Eldermet consortium (http://eldermet.ucc.ie), which has the following additional principal investigators: Ted Dinan, Daniel Falush, Gerald Fitzgerald, Tony Fitzgerald, Albert Flynn, Colin Hill, Denis O'Mahony, Fergus Shanahan, Catherine Stanton, Cillian Twomey and Douwe van Sinderen.

FUNDING

M.J.C. is funded by a fellowship from the Health Research Board of Ireland. Q.W. and J.R.C. were supported by National Research Initiative number 2008-35107-04542 from the US Department of Agriculture (USDA) National Institute of Food and Agriculture. Funding for open access charge: This paper is funded, as part of the ELDERMET project (http://eldermet.ucc.ie), by the Government of Ireland through the Department of Agriculture, Fisheries and Food, and the Health Research Board, through the Food-Health Research Initiative, 2007–2011.

Conflict of interest statement. None declared.

REFERENCES

1. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
2. Neefs,J.M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Res.*, **21**, 3025–3049.
3. Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.
4. Lane,D.J., Pace,B., Olsen,G.J., Stahl,D.A., Sogin,M.L. and Pace,N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl Acad. Sci. USA*, **82**, 6955–6959.
5. Claesson,M.J., O'Sullivan,O., Wang,Q., Nikkila,J., Marchesi,J.R., Smidt,H., de Vos,W.M., Ross,R.P. and O'Toole,P.W. (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One*, **4**, e6669.
6. Hamady,M., Walker,J.J., Harris,J.K., Gold,N.J. and Knight,R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5**, 235–237.
7. Sogin,M.L., Morrison,H.G., Huber,J.A., Welch,D.M., Huse,S.M., Neal,P.R., Arrieta,J.M. and Herndl,G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.
8. Sundquist,A., Bigdeli,S., Jalili,R., Druzin,M.L., Waller,S., Pullen,K.M., El-Sayed,Y.Y., Taslimi,M.M., Batzoglou,S. and Ronaghi,M. (2007) Bacterial flora-typing with targeted, chip-based Pyrosequencing. *BMC Microbiol.*, **7**, 108.
9. Liu,Z., DeSantis,T.Z., Andersen,G.L. and Knight,R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.*, **36**, e120.
10. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
11. Chakravorty,S., Helb,D., Burday,M., Connell,N. and Alland,D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods*, **69**, 330–339.
12. Rajilic-Stojanovic,M., Heilig,H.G., Molenaar,D., Kajander,K., Surakka,A., Smidt,H. and de Vos,W.M. (2009) Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.*, **11**, 1736–1751.
13. Schmalenberger,A., Schwieger,F. and Tebbe,C.C. (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. *Appl. Environ. Microbiol.*, **67**, 3557–3563.
14. Baker,G.C., Smith,J.J. and Cowan,D.A. (2003) Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods*, **55**, 541–555.
15. Engelbrektson,A., Kunin,V., Wrighton,K.C., Zvenigorodsky,N., Chen,F., Ochman,H. and Hugenholtz,P. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.*, **4**, 642–647.
16. Kunin,V., Engelbrektson,A., Ochman,H. and Hugenholtz,P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, **12**, 118–123.
17. Quince,C., Lanzen,A., Curtis,T.P., Davenport,R.J., Hall,N., Head,I.M., Read,L.F. and Sloan,W.T. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, **6**, 639–641.
18. Andersson,A.F., Lindberg,M., Jakobsson,H., Backhed,F., Nyren,P. and Engstrand,L. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE*, **3**, e2836.
19. Reeder,J. and Knight,R. (2009) The 'rare biosphere': a reality check. *Nat. Methods*, **6**, 636–637.
20. Wang,Y. and Qian,P.Y. (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE*, **4**, e7401.
21. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
22. Lazarevic,V., Whiteson,K., Huse,S., Hernandez,D., Farinelli,L., Osteras,M., Schrenzel,J. and Francois,P. (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J. Microbiol. Methods*, **79**, 266–271.
23. Caporaso,J.G., Lauber,C.L., Walters,W.A., Berg-Lyons,D., Lozupone,C.A., Turnbaugh,P.J., Fierer,N. and Knight,R. (2010) Microbes and Health Sackler Colloquium: global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl Acad. Sci. USA*.
24. Hummelen,R., Fernandes,A.D., Macklaim,J.M., Dickson,R.J., Changalucha,J., Gloor,G.B. and Reid,G. (2010) Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE*, **5**, e12078.
25. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W., Peplies,J. and Glockner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
26. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.
27. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
28. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for

describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
29. Garrity,G.M., Bell,J.A. and Lilburn,T.G. (2004) *Taxonomic Outline of the Procaryotes. Bergey's Manual of Systematic Bacteriology*, 2nd edn. Springer, New York.
30. Huson,D., Auch,A., Qi,J. and Schuster,S. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
31. Urich,T., Lanzen,A., Qi,J., Huson,D.H., Schleper,C. and Schuster,S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*, **3**, e2527.
32. Nawrocki,E.P. and Eddy,S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, **3**, e56.
33. Kemp,P.F. and Aller,J.Y. (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol. Ecol.*, **47**, 161–177.
34. Quail,M.A., Kozarewa,I., Smith,F., Scally,A., Stephens,P.J., Durbin,R., Swerdlow,H. and Turner,D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005–1010.
35. Dethlefsen,L., Huse,S., Sogin,M.L. and Relman,D.A. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.*, **6**, e280.
36. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
37. C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012–2018.
38. Turnbaugh,P.J., Hamady,M., Yatsunenko,T., Cantarel,B.L., Duncan,A., Ley,R.E., Sogin,M.L., Jones,W.J., Roe,B.A., Affourtit,J.P. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.
39. McKenna,P., Hoffmann,C., Minkah,N., Aye,P.P., Lackner,A., Liu,Z., Lozupone,C.A., Hamady,M., Knight,R. and Bushman,F.D. (2008) The Macaque Gut Microbiome in Health, Lentiviral Infection, and Chronic Enterocolitis. *PLoS Pathog*, **4**, e20.
40. Huber,J.A., Mark Welch,D.B., Morrison,H.G., Huse,S.M., Neal,P.R., Butterfield,D.A. and Sogin,M.L. (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.