Supplementary Material

Methods

Extraction of microbial DNA from foot skin

DNA was extracted from skin swabs taken from the feet of 3 different healthy individuals. 12 samples were taken in total. Skin swabs were collected by swabbing either the ball or heel area of the left or right foot with a rayon swab moistened in a solution of 0.15 M NaCl and 0.1% Tween 20. The swab was rubbed firmly over the skin for approximately 30 seconds. Swab heads were cut into bead beating tubes, and DNA was extracted from the swabs using the BioStic DNA extraction kit (Mo-Bio), as per the manufacturer's instructions. DNA was quantified on a Qubit with a HS-DNA assay (Life Technologies).

Preparation of short read 16S libraries for Illumina sequencing

A library of the V4 region of the 16S gene was prepared for Illumina sequencing from the microbial foot skin DNA samples using a modification of a previously published method[1]. Briefly, samples were amplified using primers based on the Caporaso et al[1] design, which were modified to include 8bp rather than 12 bp barcodes, and include a barcode on both the forward and reverse primer. The V4 region was amplified from 500 pg template DNA using 10 cycles of PCR with the modified Caporaso primers (V4_forward and V4_reverse), using different barcoded primers for each sample (Table S1). After removal of excess primer via a magnetic bead clean-up (Agencourt) samples were pooled, and subjected to a further 20 cycles of PCR to enrich for amplicons containing the Illumina adapters, using primers Illumina_E_1 and

Illumina_E_2 (Table S1). Pooling of samples during the enrichment PCR allows for an assessment of the putative recombination rate, by examining the rate of invalid barcode combinations that occur in the final paired end sequencing data. The method for each PCR reaction is described in detail below.

PCRs were carried out with a Taq core PCR kit (Qiagen), under the following conditions. For the initial 10 cycle PCR, reactions contained 1 x PCR buffer, 1 x Q solution (Qiagen), 250 μM dNTPs, 0.5 μM each of V4_forward and V4_reverse barcoded primers, 500 pg template DNA, and 1.25 U Taq DNA polymerase in a 50 μl reaction volume. Thermal cycling was carried out at 95°C for two minutes, followed by 10 cycles of 95°C for 15 seconds, 50°C for 30 seconds and 72°C for 90 seconds, followed by a final extension at 72°C for five minutes. After a magnetic bead clean-up using 0.8 volume of Agencourt beads, the cleaned PCR reactions were pooled and used as input for the second PCR reaction. This PCR contained 1 x PCR buffer, 1 x Q solution (Qiagen), 250 μM dNTPs, 0.25 μM each of Illumina_E_1 and Illumina_E_2 primers (see Table S1), 31 ul pooled PCR products from the first PCR, and 1.25 U Taq DNA polymerase in a 50 μl reaction volume. Thermal cycling was carried out at 95°C for two minutes, followed by 20 cycles of 95°C for 15 seconds, 55°C for 30 seconds and 72°C for 90 seconds, followed by a final extension at 72°C for five minutes. These PCR reactions were again cleaned via a magnetic bead clean-up as above, and run on a bioanalyser using a HS-DNA chip to confirm the amplicon size and determine the concentration.

The short read 16S libraries were sequenced using a Nano flow cell and a 500 cycle V2 kit on an Illumina MiSeq, using custom primers as described in Caporaso et al[1].

This method will be referred to as "short sequencing" and data produced with this method as "V4" data. Read pairs were merged with FLASH[2] and de-multiplexed with PhyloSift[3].

Preparation of full-length 16S libraries for Illumina sequencing with unique molecular tags

Primers for amplification of the 16S gene contained the 27F[4] or 1391R[5] bacterial primer sequences, an 8bp barcode sequence, a 10bp random tag and partial Illumina PE adapter sequences (Figure S1, Table S1). The use of a 10bp random tag on both forward and reverse primers (~1 million possible unique tags at each end, ~1 trillion combinations) allowed us to uniquely tag each 16S molecule in our pool, by modifying previously described tagging approaches[6,7]. Template DNA was subject to one cycle of annealing and extension with the forward primer (long_forward, Table S1), followed by a magnetic bead clean-up to remove excess primer, then another cycle of annealing and extension with the reverse primer (long_reverse, Table S1), followed by another magnetic bead clean-up. The first PCR carries out extension of the 16S gene from the forward primer, which uniquely tags different 16S templates in the reaction. The second PCR uses extension products from the first PCR as a template to produce molecules with unique tags at both ends. While the original 16S molecules may also act as a template in the second PCR reaction, these products will only contain an Illumina adapter at one end, and will therefore not be amplified in the enrichment PCR. The enrichment PCR (34 cycles) amplifies the tagged 16S molecule pool, using primers that are complementary to the Illumina adapter sequences at the ends of each tagged 16S molecule (primers PE_1 and PE_2, Table S1).

PCRs were carried out using the Taq PCR core kit (Qiagen), and differently barcoded primers were used for each sample. Reactions contained approximately 500 pg DNA template, 0.25 μM long_forward primer, 250 uM dNTPs, 1 x PCR buffer, 1 x Q solution, and 1.25 U Taq polymerase in a 50 μl volume. Cycle conditions were 95°C for one minute, 50°C for two minutes then 72°C for three minutes. This allows extension of the 16S gene from the forward primer, which uniquely tags the forward end of each 16S molecule in the reaction. PCR reactions were then subject to a magnetic bead clean-up using 0.6 volumes of Agencourt SPRI beads as per the manufacturers instructions, except that the DNA was eluted in 35 μl nuclease free water. The second PCR was set up as described above, except that 0.25 μM of the long_reverse primer was used, and the template was 31μl of the bead-cleaned first round annealing and extension reaction. Cycling conditions were as in the previous step: 95°C for one minute, 50°C for two minutes and 72°C for three minutes. During this second reaction, the uniquely tagged extension products from the first reaction act as the template to produce 16S molecules with unique tags on the forward and reverse ends. This was followed by another magnetic bead clean-up, as described above, and the output of this step was used as a template for the final PCR reaction. The final enrichment PCR reaction contained 0.5 μM of each PE_1 and PE_2 primers, 250 μM dNTPs, 1 x PCR buffer, 1 x Q solution, 31 μl template (from the bead clean-up) and 1.25 U Taq polymerase in a 50 ul volume. Cycling conditions were 95°C for two minutes, followed by 34 cycles of 95°C for one minute, 58°C for 30 seconds, and 72°C for two minutes, and a final extension of 72°C for five minutes. PCRs were again subject to a magnetic bead clean-up as described above, before being analysed using a high-sensitivity DNA chip on a Bioanalyser (Agilent) to determine the size and concentration.

Tagmentation of full-length 16S libraries

The uniquely tagged, full length 16S PCR amplicons were subjected to tagmentation. The tagmentation procedure utilises a transposase to simultaneously fragment the DNA while adding an adapter sequence for use on the Illumina platform. Tagmentation was carried out using the Nextera XT kit as per the manufacturer's instructions, with the exception of the PCR amplification step. Here, we split the tagmentation reaction into two, and carried out two separate PCRs at half the volume specified in the kit (where normally only one PCR is carried out). Each PCR reaction contained a combination of one of the Illumina provided Nextera XT PCR primers and one of the primers from the enrichment PCR above, so as to amplify only those fragments of interest; specifically, we combined primers PE_1 and an Illumina Index 1 primer (N706) in one PCR reaction, and PE_2 and an Illumina Index 2 primer (S504) in the second. We aimed to produce a pool of DNA fragments with either the PE_1 (forward end of the 16S amplicons) or PE_2 (reverse end of the 16S amplicons) sequences on one end, and the i7 or i5 Illumina adapters (added to an internal region of the 16S amplicon during the tagmentation reaction) at the other end, respectively. This provided a pool of fragments from across the 16S gene, which along with the full length 16S amplicons, can be paired end sequenced on the MiSeq. PCR products from the tagmentation reaction were cleaned using 0.6 volumes of Ampure SPRI beads according the manufacturer's instructions, to remove fragments smaller than 400 bp.

Sequencing of full-length and tagmented 16S libraries

The molarity of both full-length and tagmented 16S libraries was measured via a Bioanalyser High Sensitivity DNA chip. Full length 16S tagged amplicons were combined with the cleaned tagmentation products at a ratio of 1:9, loaded at an average molarity of 6 pM, and sequenced with 2 x 250 bp paired end reads, on a MiSeq Nano flow cell.

Reconstructing full length 16S sequences from tagged Illumina reads

Sequencing produces data from two kinds of fragments, those which span the entire 16S gene (end+end fragments) and those which pair one end of the 16S gene with a region in the middle of the 16S gene (end+internal fragments). Sequences from end+end fragments encode a pairing of random barcodes and sample barcodes. Sequences can be assigned to bins of original 16S progenitor molecules via the unique tags at either end of the molecule and re-assembled to provide full-length 16S sequences. Figure S2 shows an overview of the process.

To assign sequences to samples, the two 8 nt sample barcode regions are matched against the collection of known sample barcodes with up to one mismatch tolerated in each 8 nt barcode. Because internal regions of the 16S sequence might match a sample barcode, all reads with a potential sample barcode match are then screened for the presence of the proximal or distal 16S primer annealing sequence downstream from the sample barcode. Reads lacking a known sample barcode or the primer annealing sequence in one end are presumed to derive from an end+internal fragment.

Consensus random barcodes and elimination of recombinants. Due to sequencing error, the reads derived from the same template molecule may have slightly different

10nt random barcode sequences. To estimate the original 10nt random barcode sequences of tagged template molecules we apply the uclust[8,9] algorithm to identify clusters of matching random barcode sequences at >89% identity (e.g. 1 out of 10 bases mismatch), and to report the consensus sequences of these clusters. We first identify clusters of random barcodes in the end+end fragments (the clustered sequences consisting of both 10nt random barcodes, both 8 nt sample barcodes, and the first 14 nt of the 16S amplicon sequence in each read). We then identify the highest abundance cluster with each 10nt random barcode and discard any cluster containing a 10nt random barcode that was found in a different, more abundant cluster. This step aims to identify and discard combinations of random barcodes that arose due to in-vitro recombination. Recombinant forms are likely to be at lower abundance than the parental templates.

The end+end fragments may not capture all random barcodes present in a sample. The remaining random barcodes might still be used to reconstruct 16S sequences even though they can not be assigned to a sample without end+end fragment information. Therefore, we apply uclust again to identify clusters of random barcodes on each end separately, and add any new consensus sequences that were not previously found in an end+end fragment. In the present work, these clusters were not included in any further analysis

Finally, random barcodes from entire set of reads are matched against the collection of consensus sequences and the reads are grouped into clusters for later assembly.

Assembly of read clusters: Read clusters contain reads that, with high probability, originate from the same template molecule. We apply a *de novo* assembly algorithm on the read cluster to reconstruct as much of the original template molecule as possible. The reads are assembled using a version of the A5 pipeline[10] called A5-miseq that has been modified to support assembly of reads up to 500nt long and to trim out adapter sequence from reads instead of discarding reads containing adapter sequence. Only the first two stages of the A5-miseq pipeline were applied, involving adapter trimming, quality trimming, error correction, and contig assembly.

Assessment of assembled 16S long sequence quality: The accuracy of the base calls was assessed by calculating PHRED scale quality scores using samtools. Briefly, the reads present in each assembled barcode cluster were mapped back to the assembled contigs using BWA MEM. From the mapped reads, a consensus FastQ sequence was called using samtools, bcftools, and vcfutils.pl. The quality scores in the resulting FastQ file were then used for subsequent quality analysis and visualization.

Removal of chimeras in cluster assemblies: Putative chimeras are identified in end+end reads as described above, and this permits estimation of the overall recombination rate and the frequency of recombinant fragments relative to full length fragments for each cluster. However, it is not possible to identify directly end+int reads derived from a chimeric fragment using barcodes, as some of these reads will contain a molecular tag which matches an original template cluster. Erroneous signal from these reads is eliminated in two ways, both of which depend on reads derived from the recombinant form existing at lower abundance in the sequence data. First, during the initial assembly process, k-mer error correction and consensus generation

will eliminate differences in the sequence present in low abundance chimeric reads. Second, in cases where the cluster assembly contains multiple contigs, the depth of coverage of contigs is used to identify and remove contigs at much lower abundance than the dominant contigs in the cluster. For the present work we removed any contigs with an average coverage which was 10-fold lower than that of the highest abundance contig. Future work could use information derived from the end+end sequences to estimate the expected fraction of recombinant reads in a cluster and use this to aid the process of eliminating chimera-derived contigs or to identify clusters for which recombinant elimination may not be possible.

Analysis of 16S reads

Both V4 and full length data were analysed using the software package QIIME[11]. For comparison, the corresponding V4 region was extracted from the full length sequences (which we will refer to as Long-V4), and only those full length sequences that were > 1300bp in length, and therefore included the V4 region, were included in the downstream analysis. V4 sequences were initially quality filtered using the default settings, with the exception of sequence length, which was altered to remove sequences less than 240 bp and longer than 260 bp. V4 sequences were additionally assessed for the presence of chimeras using the UCHIME[12] method, both against a reference database, and using the dataset itself as the reference. Full length sequences were quality filtered using default settings and excluding sequences longer than 1400 bp. Quality filtered sequences from the V4, Full length and Long-V4 datasets were then combined, and sequences were assigned to OTUs using the closed reference picking method, which assigns sequences to pre-clustered OTUs at 97% similarity from a chimera filtered database (Greengenes)[13,14]. Taxonomy was assessed based on

membership to the database of pre-clustered OTUs, using the *summarize_taxa.py* script.

In order to demonstrate whether increased phylogenetic resolution was possible with full length sequences, phylogenetic trees were constructed from sequences greater than 1300 bp and compared to trees constructed from the V4 (Long-V4) and V1-V4 (Long-V1-4) regions of the same data set. Secondary structure aware alignments of the Long-Read sequences were computed with the Infernal software package, and the portions of the alignment corresponding to the V1-V4 and the V4 regions were extracted to obtain corresponding Long-V1-4 and Long-V4 alignments. Phylogenies were then inferred with FastTree2. The number of resolved branches in each phylogeny was reported, and one-sided Kolmolgorov-Smirnov tests were carried out to check whether the clade support values were higher in the Long-Read relative to Long-V1-4 and Long-V4.

Results

Full length 16S sequences generated by molecular tagging

Clustering of end+end reads resulted in 5085 clusters. Of these, 2265 (44.6%) were deemed to be putative recombinant clusters, with parental templates on average 29 times more abundant than putative recombinants (Figure S3). Putative recombinant end+end sequences represented 4378 of the total 42715 sequences in the end+end read pool, indicating an average recombination rate of 10.2 % among all samples. After binning and assembly of end+end and end+internal read clusters, 2304 16S sequences were assembled from 558,053 Illumina read pairs. Sequence lengths

ranged from 449 to 1372 bp (full length), and 70% were greater than 1300 bp (Figure S4).

Assembled sequences had consistently high quality scores across their length, with average estimated PHRED quality scores at each position ranging from 54.0 – 89.5 (median 68.0) (Figure 2a). This indicates base calling accuracies of greater than 99.999% at each position of the assembled 16S sequences. Quality scores were higher at either end of the 16S sequences, due to the increased coverage of these regions as a result of every end+internal sequence covering the same region of one or the other end of the 16S molecule (Figure 2b).

Short sequencing of the 16S V4 region

A total of 296,864 paired end V4 sequences were generated from 12 foot skin samples. Of these sequences, 11,240 could not be assigned to a sample due to invalid forward and reverse barcode combinations (e.g. combinations which were never assigned to a sample), indicating an *in-vitro* recombination rate of at least 3.8%. These sequences were removed from the dataset. We note that *in-vitro* recombination could also create barcode combinations that would match a valid sample and therefore be undetectable recombination events. In contrast, when attempting to detect recombination products using the chimera detection software UCHIME (as implemented in QIIME), only 0.05% of the sequences were flagged as chimeric when compared against a reference database (SILVA), and 0.2% when using the dataset itself as the reference. This highlights the difficulties of using software alone to detect recombination products from PCR in the absence of sample barcode and molecular tag information. Sequences that were flagged as chimeric using UCHIME,

which had not been identified as chimeric based on sample barcode combinations (as described above) were also removed from the dataset.

Assembled full length 16S sequences produce data consistent with short read sequencing.

Taxonomy, as assessed in QIIME by membership to the database of pre-clustered OTUs, was similar to previous reports for skin communities, dominated by *Firmicutes*, *Actinobacteria*, and *Proteobacteria*. Full length and Long_V4 OTUs showed the same broad taxonomic distribution as the V4 sequence data (Figure 2B). There was a small decrease in the representation of *Firmicutes*, and an increase in the representation of *Actinobacteria* and *Proteobacteria* (Figure 2c), however these differences were not significant (two tailed t-test, $p > 0.05$). Similar taxonomic assignments between the different sequencing methods were also observed at the level of genera, with communities dominated by *Staphylococcus*, followed by *Corynebacterium*, *Enhydrobacter* and *Acinetobacter*. The *Corynebacterium* genus had an increased representation in the full length data set as compared to the V4 data, which likely accounts for the observed difference in representation for the *Actinobacteria* phyla, but as above, this difference was not significant (two tailed t-test, $p > 0.05$).

Comparison at the OTU level: Of the OTUs clustered at 97% similarity from the full length sequence data, an average of 22.7 % (±15.6) were also found in matched sample V4 data that was clustered in the same way. This disparity is likely to be due to comparing OTUs of sequences of different lengths, and the way in which OTUs are defined in QIIME. Sequences are assigned to OTUs by the best match against a

database of representative sequences which have been pre-clustered into OTUs[15]. Presumably, full length or long sequences from the database were used to cluster OTUs, and clusters that are 97% similar across the full 16S gene may not be 97% similar in the V4 region only, since different regions of the 16S gene evolve at different rates[16]. We therefore analysed OTUs clustered from the V4 region of the full length sequences (Long-V4 sequences) to assess whether we had captured similar OTUs with the V4 and full length sequencing methods. In this case 88.4 % (±15.9) of Long-V4 OTUs were shared with the matched sample V4 OTUs (Table S2). Although fewer sequences were present in the full length data set, yielding many fewer OTUs overall, the data indicates that the newly developed method gives broadly congruent community profiles with respect to taxonomy and OTU clustering.

Assessment of increased phylogenetic resolution using full length sequencing vs shorter fragments of the 16S gene

Analysis of phylogenetic trees constructed from full length sequences, and trees from the corresponding V1-V4 and V4 regions of the same sequence set showed that full length sequences resolved more of the possible branches with higher confidence._Full length sequences resolved 2954 of a possible 3179 branches, compared to 2686 for the V1-V4 region, and 2114 for the V4 region. Kolmolgorov-Smirnov tests rejected the hypothesis that V4 has higher support ($p= 0.003$), and that phylogeny on V1-V4 yields higher support values ($p=1.67x10^{-13}$). Figure S5 shows the distribution of confidence values for nodes in the full length, Long-V1-4 and Long-V4 phylogenies.

Assessment of bias reduction using unique molecular tags

The use of molecular tagging has previously been shown to reduce the effect of PCR bias in RNA-seq data, for better quantitative assessment of sequences from the original samples[17]. Assuming that each uniquely tagged 16S molecule from our skin samples should have been present at the same abundance as all other uniquely tagged molecules (i.e. 1 copy of each), and that unbiased amplification would result in an equal abundance of each cluster, we can estimate the amount of biased amplification that occurred during PCR by comparing the differences in the abundance of end+end sequence clusters. The average abundance was calculated from all clusters, and the relative mean error was 2.08, or 1.81 if singleton clusters (possible recombinants) were excluded. This indicates a standard deviation of approximately 2 times the average across the dataset under the particular amplification conditions used here. Figure S5 shows the distribution of the estimated amplification bias, which ranges from 0.06 to ~ 32 times the average cluster abundance. This potential bias is eliminated by considering each assembled 16S sequence cluster as having a count of 1.

References

1 Caporaso, J. G. *et al. The ISME journal* **6**, 1621-1624, doi:10.1038/ismej.2012.8 (2012).

2 Magoc, T. & Salzberg, S. L. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

3 Darling, A. E. *et al. PeerJ* **2**, e243, doi:10.7717/peerj.243 (2014).

4 Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. *Journal of bacteriology* **173**, 697-703 (1991).

5 Turner, S., Pryer, K. M., Miao, V. P. & Palmer, J. D. *The Journal of eukaryotic microbiology* **46**, 327-338 (1999).

6 Faith, J. J. *et al. Science* **341**, 1237439, doi:10.1126/science.1237439 (2013).

7 Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. *Nature methods* **10**, 999-1002, doi:10.1038/nmeth.2634 (2013).

8 Edgar, R. C. *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

9 Edgar, R. C. *Nature methods* **10**, 996-998, doi:10.1038/nmeth.2604 (2013).
10 Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. *PloS one* **7**, e42304, doi:10.1371/journal.pone.0042304 (2012).
11 Caporaso, J. G. *et al. Nature methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).
12 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. *Bioinformatics* **27**, 2194-2200, doi:10.1093/bioinformatics/btr381 (2011).
13 DeSantis, T. Z. *et al. Applied and environmental microbiology* **72**, 5069-5072, doi:10.1128/AEM.03006-05 (2006).
14 McDonald, D. *et al. The ISME journal* **6**, 610-618, doi:10.1038/ismej.2011.139 (2012).
15 Rideout, J. *et al. PeerJ PrePrints* (2014).
16 Schloss, P. D. *PLoS computational biology* **6**, e1000844, doi:10.1371/journal.pcbi.1000844 (2010).
17 Islam, S. *et al. Nature methods* **11**, 163-166, doi:10.1038/nmeth.2772 (2014).

Supplementary Figures

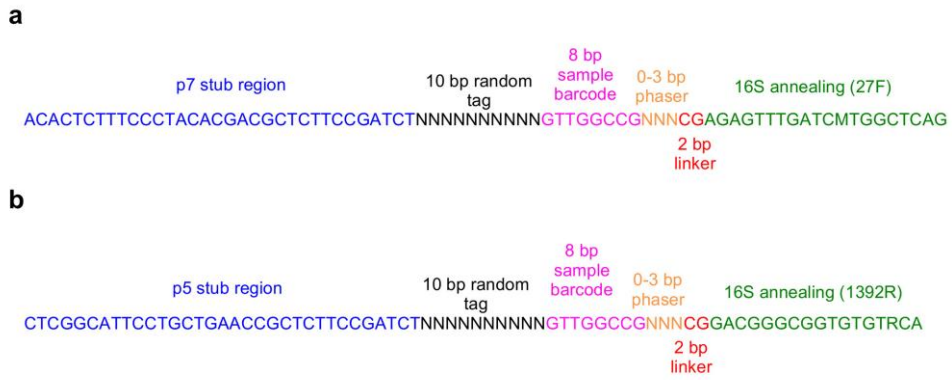

Figure S1: Schematic of primers used for molecular tagging of 16S template molecules. a) long_forward and b) long_reverse. Stub regions correspond to Illumina adaptors for clustering on the MiSeq, and 0-3 bp phasers are included to ensure diversity between barcoded samples during sequencing. 25 different barcodes were designed for up to 625 different sample barcode combinations, which are listed in Table S1.

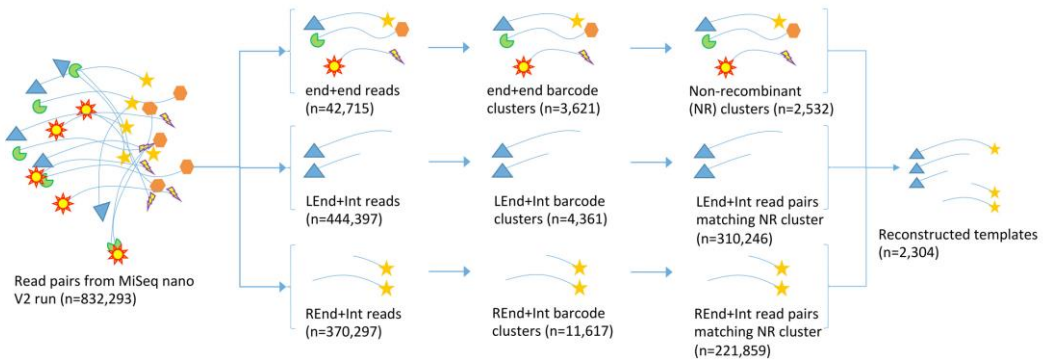

Figure S2: Schematic demonstrating the processing of read pairs from the MiSeq to reconstruct Long-Read 16S sequences. Read pairs are placed into groups of end + end sequences, or end + internal sequences. End + end sequences are clustered into groups containing the same combination of random molecular tags from either end, and putative recombinant clusters are removed (identified as having one or two molecular tags from a separate, more abundant cluster). End + internal sequences are assigned to clusters based on their unique molecular tags, and each cluster is used to generate an assembly of the full length sequence.

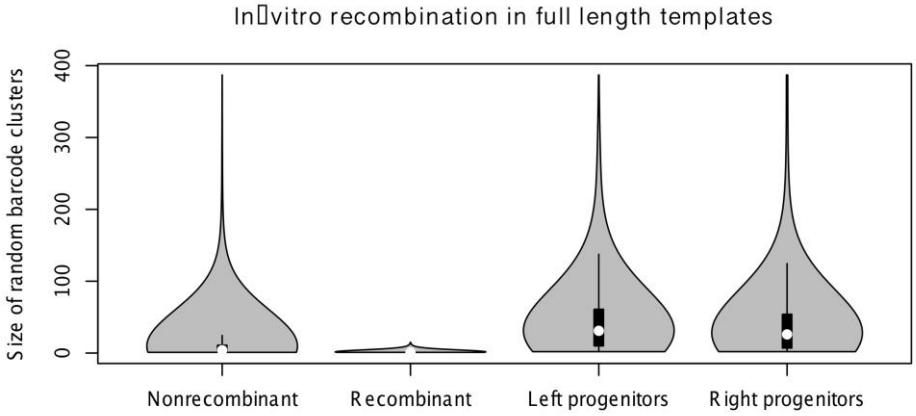

Figure S3: Abundance of putative recombinants. Violin plot showing the abundance of barcode clusters identified as putatively recombinant (left), along with abundances of the progenitor molecules producing recombinant forms. Parental templates were on average 29 times more abundant than the putatively recombinant forms. Median values are indicated by white dotes, and the interquartile range by black boxes.

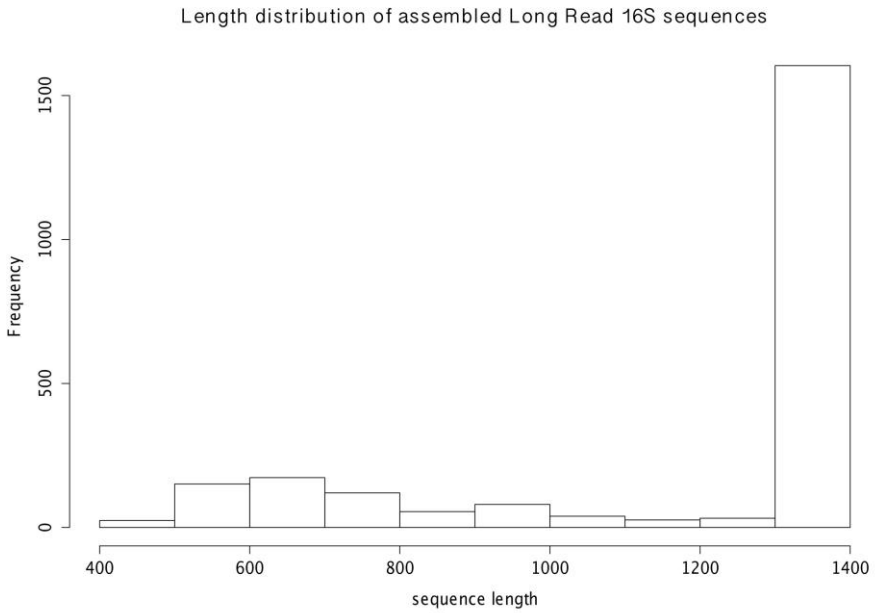

Figure S4: The length distribution of assembled Long-Read 16S sequences. Sequence length ranged from 400bp to 1378bp, corresponding to a full length amplicon. 70% of the assembled sequences are >1300bp in length.

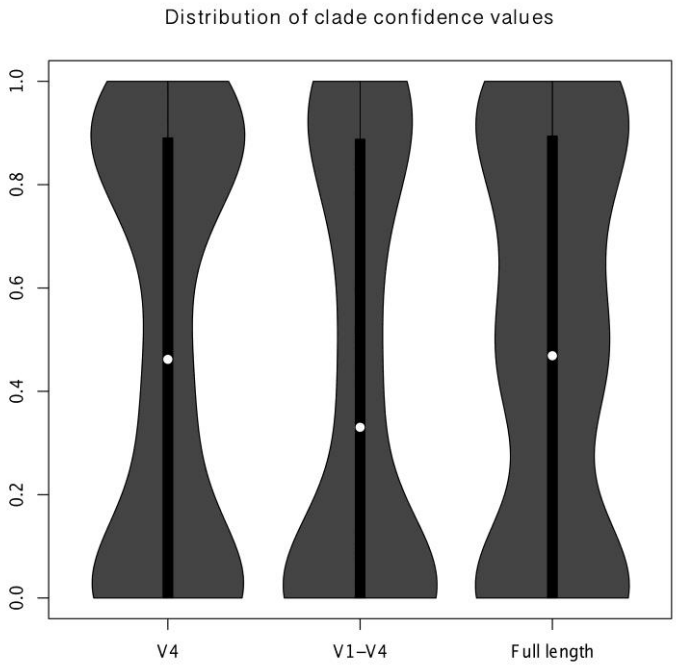

Figure S5: Confidence value distributions for phylogenies constructed from Long-Read sequences, and the corresponding V1-4 and V4 regions. The V4 region resolved less branches overall and with slightly lower confidence than the Long-Read sequences, while the V1-V4 resolved more branches than the V4 region, the confidence values were significantly lower. Data is plotted as a violin plot, with median values indicated by white dots, and the interquartile range by black boxes.

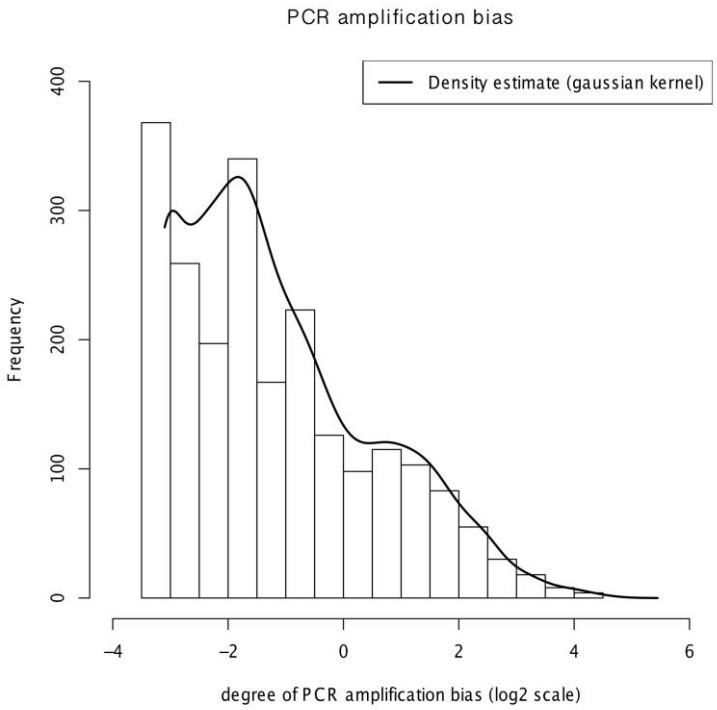

Figure S6: Distribution of the estimated degree of PCR amplification bias. Estimates of bias were calculated from the deviation of each end + end sequence cluster from the mean end + end sequence cluster abundance.