

1 **Resolving microbial microdiversity with high accuracy, full length 16S rRNA Illumina**
2 **sequencing**

3

4 Catherine M. Burke^{1*} and Aaron E. Darling^{1*}

5 1. The ithree institute, University of Technology Sydney, NSW, Australia

6

7 * Corresponding author: Catherine Burke, email: catherine.burke@uts.edu.au

8 * Corresponding author: Aaron Darling, email: aaron.darling@uts.edu.au

9

10 Running title: High precision full length Illumina 16S sequencing

11

12

13 **ABSTRACT**

14 We describe a method for sequencing full-length 16S rRNA gene amplicons using the high
15 throughput Illumina MiSeq platform. The resulting sequences have about 100-fold higher accuracy
16 than standard Illumina reads and are chimera filtered using information from a single molecule dual
17 tagging scheme that boosts the signal available for chimera detection. We demonstrate that the data
18 provides fine scale phylogenetic resolution not available from Illumina amplicon methods targeting
19 smaller variable regions of the 16S rRNA gene.

20

21 INTRODUCTION

22 Amplifying and sequencing 16S rRNA genes from microbial communities has become a standard
23 technique to survey and compare communities across space, time and environments. High-
24 throughput sequencing methods have made bacterial community profiling routine and affordable,
25 but at the expense of read length, with most platforms covering 250 to 600 bp of the ~1500 bp 16S
26 rRNA gene. Depending on the region sequenced, these shorter fragments give variable taxonomic
27 and phylogenetic resolution¹⁻³ and fails to resolve differences outside the targeted region, which
28 may be ecologically relevant⁴⁻⁶.

29

30 Interpretation of 16S amplicon sequencing data is confounded by PCR and sequencing artefacts
31 including chimeras, biased amplification and sequencing errors. Some of these artefacts can be
32 removed computationally⁷, but nevertheless lead to errors that artificially inflate diversity
33 estimates⁸⁻¹⁰ and mislead analysis.

34

35 RESULTS

36 We present a method to sequence full-length 16S amplicons using Illumina technology. The
37 technique incorporates randomized molecular tags on both ends of individual 16S template
38 molecules prior to PCR amplification. Copies of the templates are fragmented and sequenced and
39 the dual tag information is used to accurately re-assemble full length 16S gene sequences. In
40 addition to generating longer sequences, this method offers large reductions in base calling errors,
41 reduced amplification bias¹¹ (see Supplementary materials), and a molecular signal supporting the
42 identification and removal of artefacts generated by *in-vitro* recombination (chimeras). We
43 evaluate this method by application to the skin microbiome and demonstrate improved phylogenetic
44 resolution relative to short read sequencing of the V4 or the V1-4 regions. An overview of the
45 method is shown in Figure 1.

46

47 Previous approaches for molecular tagging of the 16S rRNA gene have targeted individual variable
48 regions, rather than sequencing of the whole gene, have incorporated tags at only one end of the
49 amplicon⁸, or tags too short¹⁰ to permit the assembly and chimera filtering techniques proposed
50 here. We tagged full-length individual 16S rRNA gene template molecules with unique molecular
51 tags at both ends (Figure 1a) in DNA samples from foot skin swabs. This was achieved through
52 primers incorporating a randomized 10 nucleotide sequence in addition to the gene priming region,
53 sample specific barcodes, and adapters for Illumina sequencing (Figure S1). Tagged 16S rRNA
54 gene templates were then amplified via PCR to give a mixed pool of 16S amplicons, where those
55 molecules with the same combination of unique tags at either end were highly likely to have
56 originated from the same template. The 10 base pair random sequence allows for $4^{10} \approx 1$ million
57 possible unique tags at each end of the tagged molecules, and ~ 1 trillion unique tag combinations.

58

59 After PCR amplification, aliquots of the 16S libraries were fragmented and prepared for Illumina
60 sequencing using the standard tagmentation approach (Nextera XT, Illumina)¹². For the purpose of
61 full length 16S sequencing, the PCR step was altered to selectively amplify only fragments
62 containing one or the other of the original tagged 16S template molecule ends, coupled with a
63 Nextera adapter introduced inside the original template via tagmentation. Sequencing of both the
64 full length and the tagmented amplicon pools was then carried out on a single MiSeq Nano flow cell
65 using standard primers, yielding 832,293 read pairs.

66

67 Paired end sequences from the full length amplicons, referred to as end + end sequences (n=42715),
68 were used to determine which pairs of unique molecular tags from either end of the amplicons were
69 from the same progenitor molecule. A clustering algorithm was used to define groups of frequently
70 occurring molecular tags in the presence of sequencing error, resulting in 5085 clusters of end + end
71 sequences. Putative *in vitro* recombination products were identified as low abundance clusters
72 containing a molecular tag on one or both ends that was also observed in a higher abundance cluster

73 (see online methods). Putative recombinant clusters were on average 29 times less abundant than
74 the non-recombinants, and represented 4378 of the 42715 end + end sequences, indicating a
75 recombination rate of 10.2%. These recombinant forms are a common problem in amplicon
76 sequencing¹³, and are difficult to detect in the absence of molecular tags (as demonstrated on short
77 read V4 sequences, see Supplementary results). Their removal ensures a more accurate
78 representation of the true diversity captured from each sample.

79

80 Reads from the tagged amplicon pools were then assigned to previously defined clusters of
81 molecular tag pairs, and assemblies (see online methods) were performed on read clusters to yield
82 consensus 16S sequences, ranging in length from 400 bp to 1378 bp (full length). In a single MiSeq
83 Nano run of 832,293 read pairs, we assembled 2304 sequences, 70% of which were greater than
84 1300bp (Figure S5). Assuming linear scaling, the method would yield up to 80,000 full length 16S
85 sequences on a 600 cycle MiSeq v3 kit, while a HiSeq 2500 might generate more than 480,000 full
86 length 16S sequences in a single “rapid run” lane.

87

88 For comparison, the same samples were subject to sequencing of the V4 region (250 bp) of the 16S
89 gene with an adaptation of a widely used protocol¹⁴ (see online methods), using a MiSeq Nano flow
90 cell with 150 bp paired end reads. Figure 2 shows the results obtained using the V4 and Long-Read
91 protocols. Higher average quality scores were obtained for the new sequencing protocol
92 (68.6 ± 17.2 , PHRED scale, estimated by bcftools) as compared to the V4 sequencing (37.4 ± 1.3).
93 The 100-fold reduction in average error rate resulted from the increased depth of coverage of
94 individual 16S molecules (Figure 2a-b).

95

96 Taxonomic assignments were compared between the V4 sequencing data and the V4 region of the
97 full length data (long_{V4}), to exclude differences in taxonomic assignment due to sequence length.
98 Sequences were analysed using QIIME¹⁵ for OTU clustering and taxonomic assignment (see

99 Supplementary methods). Similar taxonomic assignments were obtained at broad taxonomic
100 classification levels (Phylum and Genus, Figure 2c) for both the V4 and long_{V4} sequences.

101

102 In order to evaluate how the new protocol influences phylogenetic resolution, we compared
103 phylogenies constructed from sequences greater than 1300 bp to those constructed from the V4
104 (long_{V4}) and V1-V4 (long_{V1-4}) regions of the same data set. Phylogenies on full-length sequences
105 were better resolved (2954 of a possible 3179 branches compared to 2686 for the V1-V4 region,
106 and 2114 for the V4 region) with stronger support (Figure S5). While molecular tagging has
107 previously been demonstrated to eliminate false microdiversity^{8,10}, our data indicates it can also
108 capture additional true microdiversity when applied to longer amplicons.

109

110 CONCLUSION

111

112 We have demonstrated accurate reconstruction of full length 16S amplicon sequences from high
113 throughput Illumina sequencing data. Although a protocol to sequence full length 16S amplicons on
114 the Pacific Biosciences instrument has been demonstrated¹⁶, it can not at present match the
115 throughput and cost efficiency of current Illumina platforms. We estimate nearly half a million full
116 length 16S sequences could be generated per HiSeq 2500 lane. We note that the method of dual
117 molecular tagging could be applied to any sequencing platform and any amplicon target to enhance
118 chimera removal and reduce amplification bias and base calling error. This method, in conjunction
119 with new algorithms¹⁷, will enable a finer understanding of population dynamics in microbial
120 ecosystems.

121

122 ACCESSIONS

123 Raw read data is available from SRA accession SRX655489. Assembled 16S sequences have been
124 provided as a Supplementary data set, Long-Read.fasta

125

126 Software automating the process of sample demultiplexing, barcode clustering, and amplicon
127 assembly is available from <http://github.com/koadman/longas> (to be made public).

128

129 ACKNOWLEDGMENTS

130 We thank Josh Quick for advice on configuring the MiSeq to cluster long fragments, Paul Worden
131 for assistance operating the MiSeq instrument, and Torsten Thomas for helpful discussions and
132 suggestions during manuscript preparation.

133

134 REFERENCES

- 135 1 Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R. P. *et al.*
136 *Nucleic acids research* **38**, e200, doi:10.1093/nar/gkq873 (2010).
- 137 2 Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A. & De Vos, P. *PloS one* **8**, e71360,
138 doi:10.1371/journal.pone.0071360 (2013).
- 139 3 Schloss, P. D. *PLoS computational biology* **6**, e1000844, doi:10.1371/journal.pcbi.1000844
140 (2010).
- 141 4 Denef, V. J., Kalnejais, L. H., Mueller, R. S., Wilmes, P., Baker, B. J., Thomas, B. C. *et al.*
142 *Proceedings of the National Academy of Sciences of the United States of America* **107**,
143 2383-2390, doi:10.1073/pnas.0907041107 (2010).
- 144 5 Fitz-Gibbon, S., Tomida, S., Chiu, B. H., Nguyen, L., Du, C., Liu, M. *et al. J Invest*
145 *Dermatol* **133**, 2152-2160, doi:Doi 10.1038/Jid.2013.21 (2013).
- 146 6 Moore, L. R., Rocap, G. & Chisholm, S. W. *Nature* **393**, 464-467, doi:10.1038/30965
147 (1998).
- 148 7 Schloss, P. D., Gevers, D. & Westcott, S. L. *PloS one* **6**, e27310,
149 doi:10.1371/journal.pone.0027310 (2011).
- 150 8 Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L.
151 *et al. Science* **341**, 1237439, doi:10.1126/science.1237439 (2013).
- 152 9 Kunin, V., Engelbrekton, A., Ochman, H. & Hugenholtz, P. *Environmental microbiology*
153 **12**, 118-123, doi:10.1111/j.1462-2920.2009.02051.x (2010).
- 154 10 Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. *Nature*
155 *methods* **10**, 999-1002, doi:10.1038/nmeth.2634 (2013).
- 156 11 Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M. *et al. Nature methods*
157 **11**, 163-166, doi:10.1038/nmeth.2772 (2014).
- 158 12 Adey, A., Morrison, H. G., Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B. *et al.*
159 *Genome biology* **11**, R119, doi:10.1186/gb-2010-11-12-r119 (2010).
- 160 13 Ahn, J. H., Kim, B. Y., Song, J. & Weon, H. Y. *Journal of microbiology* **50**, 1071-1074,
161 doi:10.1007/s12275-012-2642-z (2012).
- 162 14 Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., *et al.*
163 *The ISME journal* **6**, 1621-1624, doi:10.1038/ismej.2012.8 (2012).
- 164 15 Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K.
165 *et al. Nature methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).
- 166 16 Fichot, E. B. & Norman, R. S. *Microbiome* **1**, 10, doi:10.1186/2049-2618-1-10 (2013).

167 17 Tikhonov, M., Leach, R. W. & Wingreen, N. S. *The ISME journal*,
 168 doi:10.1038/ismej.2014.117 (2014).
 169

170 **Figures**

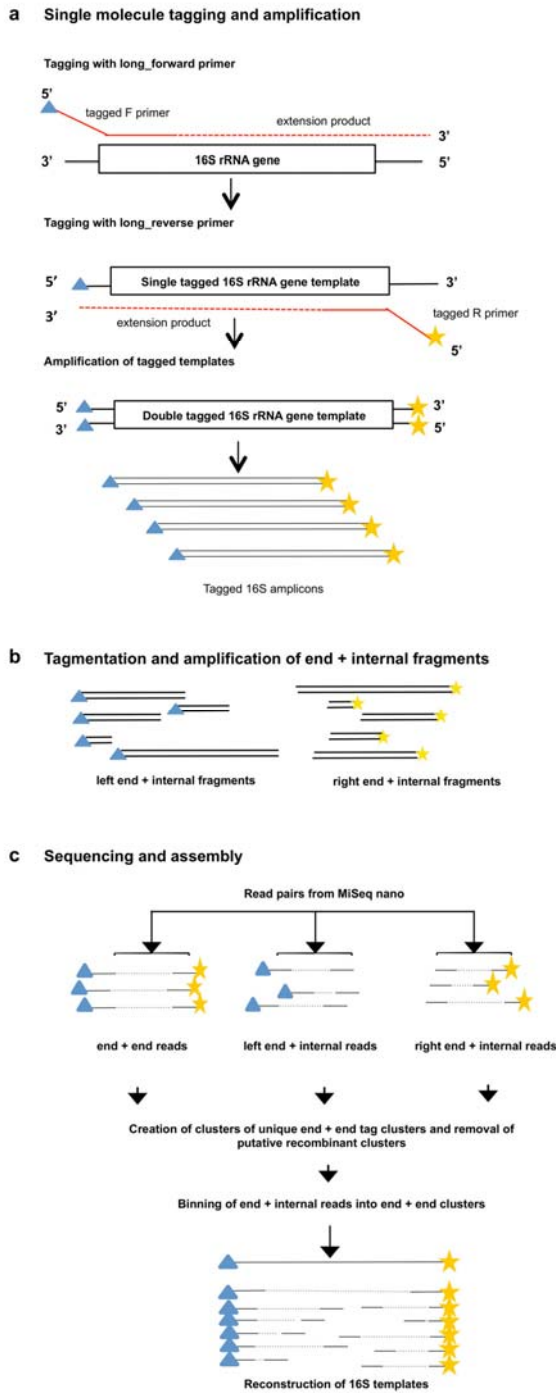


Figure 1. Overview of the Long-Read

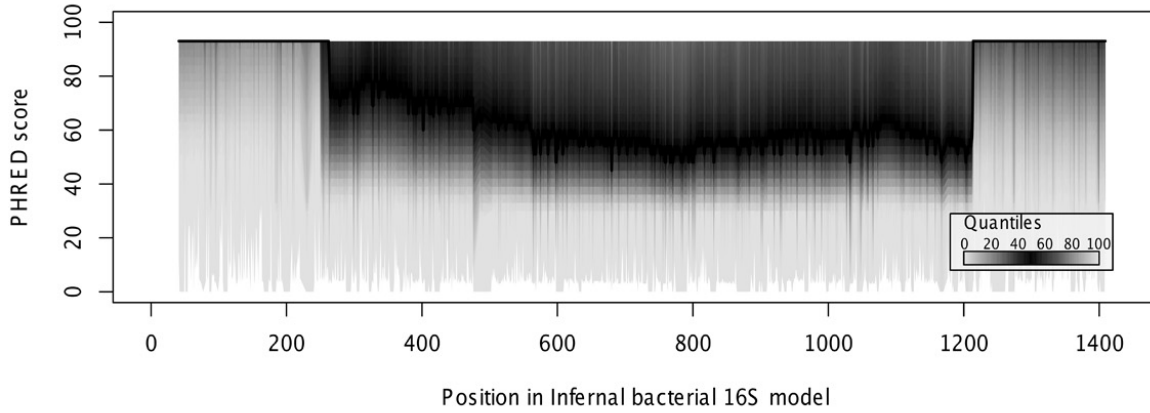
method.

a) 16S rRNA gene template molecules are tagged with unique tags via two single rounds of annealing and extension with tagged forward and reverse primers (see Figure S1), that also contain Illumina adapter sequences. After removal of tagged primers, tagged templates are amplified via PCR using primers complementary to the adapter sequences. Libraries from one or more samples can then be pooled and sequenced on the MiSeq. Blue triangles and yellow stars indicate random tags of 10bp. b) Full-length 16S amplicon Illumina libraries are tagged using the standard Nextera method, and two pools of products are amplified, which contain either the left end of the tagged amplicons and an internal region, or the right end of the amplicon and an internal region. This procedure adds Nextera adapters for sequencing at the internal end of the fragments. c) Both full

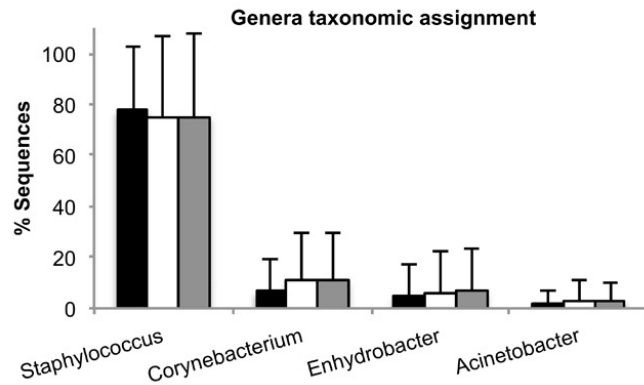
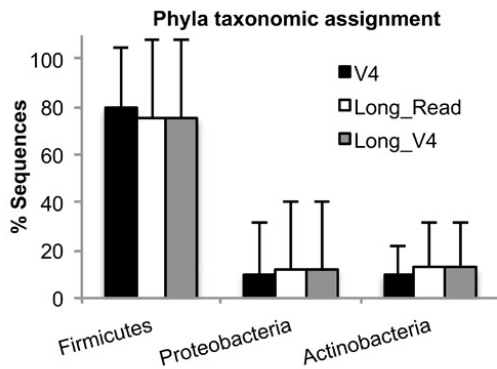
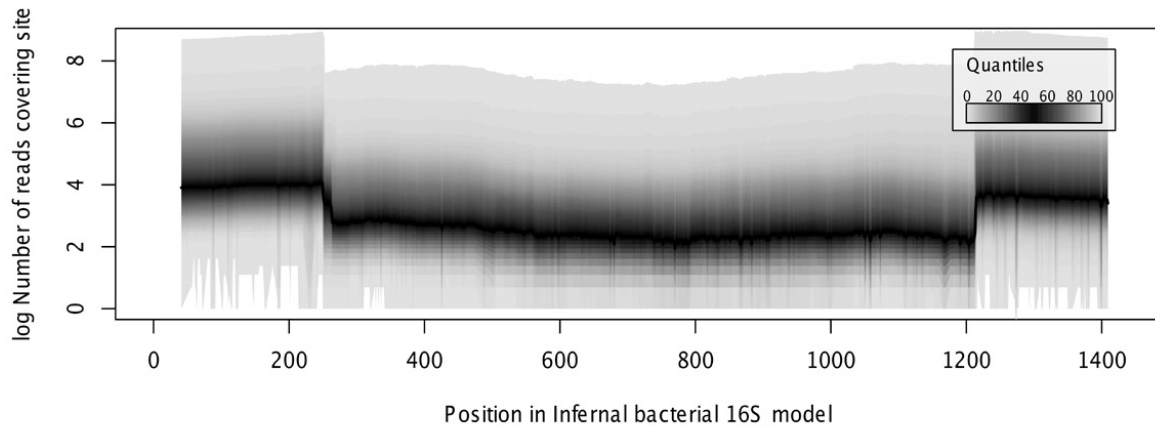
193 length and tagmented libraries are paired end sequenced, and the unique molecular tags are used to

194 computationally cluster sequences from the same progenitor 16S rRNA gene molecule for assembly
195 of full length sequences.
196

Base call qualities in 16S sequence



Depth of coverage in 16S microassemblies



197

198 **Figure 2. Results for Long-Read and V4 16S sequencing.** a) Quality and b) coverage plots
199 showing median PHRED quality scores and associated coverage across the length of the Long-Read
200 16S sequences. Higher qualities for the long-read sequences are associated with higher coverage,
201 which is particularly apparent at each end of the amplicons (up to 200 and beyond 1200 bp), which
202 were associated with one read from every read pair in the data set. c) Taxonomic assignments at the
203 phylum and genus level of OTUs for V4, Long-Read and Long-V4 sequences (V4 region extracted
204 from the Long-Read sequences). Similar taxonomic assignments are obtained across the V4 and
205 Long-Read datasets. Only taxa comprising an average of more than 1% of sequences per sample
206 are displayed. Bars represent the average value across 12 samples for each sequencing method, and
207 error bars are the standard deviation.
208