One-Step Making of DNA Dumbbells Allows Tunneling of Short-Read Libraries into a Long-Read Format

Ya-Hui Chang[1], Hong-Hsing Liu[1,2,*]

[1] Institute of Molecular and Genomic Medicine, National Health Research Institutes, Zhunan 35053, Taiwan

[2] Pediatrics, En Chu Kong Hospital, Sanxia 23702, Taiwan

*Correspondence: hhliu@nhri.edu.tw; honghsing.liu@gmail.com

Abstract

DNA dumbbells have many applications. One of the useful features is to have DNA amplified in rolling circles. Repeated iterations allow long and accurate sequencing from single molecules. We have developed a highly efficient and simple method for 'tunneling' short-read Illumina sequencing libraries into long-read dumbbells. A 5'-exonuclease used in Gibson Assembly converted fixed Illumina P5/P7 sequences into sticky ends. Self-looping oligonucleotides with complementary 3' overhangs were ligated to form dumbbells. The reaction proceeded in one pot and modified only end nucleotides of input libraries. Resultant dumbbells were ready for Phi29-dependent rolling circle amplifications. Demonstrative readouts of a standard microbial community on the platform of Pacific Biotechnology confirmed successful tunneling of the original P5/P7-ended sequencing libraries.

Keywords: Dumbbell DNA, Rolling Circle Amplification, Next-Generation Sequencing

Introduction

DNA dumbbells are a variant of circular DNA. With topological equivalence to circles, DNA dumbbells are highly resistant to exonucleases and can serve as stable vectors to regulate gene expressions[1]. Beyond stability, these special forms of DNA can be used as iterative templates to amplify DNA in rolling circles[2]. Many modern sequencers have taken advantage of this feature to use DNA dumbbells or circles as sequencing templates. Pacific Biotechnology (PacBio), for example, can derive highly accurate and long sequences *via* alignments of iterative 'subreads'[3]. MGI Tech instead uses circular DNA to amplify sequencing signals *via* so-called 'DNA nanoballs'[4].

Advances in parallel sequencing have made possible that a complete human genome can be analyzed within hours[5]. Decreased costs and increased throughputs

have accelerated application of these sequencing technology in all fields of life science. However, the need to transform sequencing subjects into suitable library formats remains a key pragmatic constraint. For example, short-read libraries for Illumina are linear DNA with fixed P5/P7 sequences on either end[6]. Long-read libraries for PacBio are instead dumbbell-shaped with fixed loop sequences[7]. These libraries require different workflows to prepare. Each sequencing platform has its own advantages and disadvantages[8]. Illumina platforms are highly accurate but suffer from short reads, namely 150bp paired-end on most recent hardware. PacBio platforms instead offer long reads well over 10kbp, but accuracy is only achieved after aligning independent single-pass repeats to gain consensus 'HiFi' sequences[3]. For some libraries, such as those that include full-length heavy chain sequences from B lymphocytes, neither platform can be readily used. These libraries are typically longer than what Illumina platforms can comfortably sequence, but there are few established workflows to format these libraries for PacBio platforms. A potential solution to the dilemma is to unify the protocols while retaining a 'tunneling' option. We achieved this option with an innovative one-step making of DNA dumbbells.

Materials and Methods

Preparation of 16S rDNA

16S rDNA of variable lengths from the DH5α *E. coli* strain were PCR amplified using the following thermocycle: 95°C 3 minutes, 30 cycles of 98°C 20 seconds + 60°C 15 seconds + 72°C 2 minutes, and 72°C 7 minutes. Each 50 µl reaction contained either a pellet from 10 µl of 5-hour culture or 10 ng template DNA, 0.2 µM each primer, and KAPA HiFi HotStart ReadyMix (Roche). A full-length founder rDNA of 1.6 kbp was prepared with the primer pair 5'-AAGCAGTGGTATCAACGCAGAGT AGRGTTYGATYMTGGCTCAG-3' and 5'-CAGACGTGTGCTCTTCCGATCT RGYTACCTTGTTACGACTT-3'. P5 and P7-flanking derivatives of lengths varying by factors of two were subsequently amplified from the template by a common primer 5'-GGCGACCACCGAGATCTACAC AGRGTTYGATYMTGGCTCAG-3' coupled with one of the primers 5'-GCAGAAGACGGCATACGAGAT RGYTACCTTGTTACGACTT-3', 5'-GCAGAAGACGGCATACGAGAT CCAAGTCGACATCGTTTACGG-3', 5'-GCAGAAGACGGCATACGAGAT CCTTCCTCCCCGCTGAAAGTA-3', 5'-GCAGAAGACGGCATACGAGAT GCACATCCGATGGCAAGAGG-3', and 5'-GCAGAAGACGGCATACGAGAT ACATTACTCACCCGTCCGCCACT-3' for each template.

Dumbbell formation

16S rDNA inputs reacted with two self-looping oligonucleotides (SLOs) in

77 NEBuilder® HiFi DNA Assembly Master Mix (New England BioLabs®) at 45°C for 2-3
78 hours. The 5'-ends of SLOs were phosphorylated and protected by five consecutive
79 phosphorothioate bonds (*). The 3'-ends of SLOs were compatible with the
80 corresponding end sequences of indicated inputs. Products in Fig. 1 were formed
81 from the SLO pair Pho5'-G*G*C*C*A*GCAGC AGAGGAGGAC GCTGCTGGCC ATATG
82 AATGATACGGCGACCACCGAGATCTACAC-3' and Pho5'-G*A*A*G*T*GCGCTG
83 TAAGTATTAC CAGCGCACTTC ATAT CAAGCAGAAGACGGCATACGAGAT-3', while those in
84 Fig. 2 were formed from the SLO pair Pho5'-G*G*C*C*A*GCAGC AGAGGAGGAC
85 GCTGCTGGCC AAGCAGTGGTATCAACGCAGAGT-3' and Pho5'-G*A*A*G*T*GCGCTG
86 TAAGTATTAC CAGCGCACTTCT ATAATGG CAGACGTGTGCTCTTCCGATC-3'.
87
88 **Efficiency calculation**
89 Successfully formed dumbbells resist exonuclease digestion, but DNA inputs with
90 open ends do not. To calculate efficiency, one half of each cleanup prepared using
91 MinElute PCR Purification Kit (QIAGEN) was treated with a cocktail of exonucleases
92 comprising 0.2 U/ µL T7 exonuclease, 0.2 U/ µL truncated exonuclease VIII, 0.2 U/ µL
93 exonuclease I, and 1 U/ µL exonuclease III (New England BioLabs®), while the other
94 half of each cleanup was not treated. After 30 minutes at 25°C followed by 30
95 minutes at 37°C, remaining DNA in both halves was quantified using a DNA Qubit™
96 dsDNA HS Assay Kit (Invitrogen™). Efficiency of dumbbell formation was derived
97 accordingly.
98
99 **PCR verification**
100 A common primer, 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
101 GACTACHVGGGTATCTAATCC-3' or 16S-V4-R, was paired with loop-hybridizing primer
102 SL (5'-GTCCATCCTCGTCCTCCTC-3') or SL10Ta (5'-AGCAGC GTCCTCCTCT GCTG-3') to
103 verify dumbbell formation, and 16S-V3-F (5'-
104 ACACTCTTTCCCTACACGACGCTCTTCCGATCT CCTACGGGNGGCWGCAG-3') to control
105 inputs. PCR reactions were performed in Q5 HiFi polymerase ready mix (New England
106 BioLabs®) using the following thermocycle: 98°C 30 seconds; 5 cycles of 98°C 10
107 seconds, 45°C (SL) or 62°C (SL10Ta or 16S-V3-F) 3 minutes; 25 cycles of 98°C 10
108 seconds, 54°C (SL) or 62°C (SL10Ta or 16S-V3-F) 30 seconds, 72°C 40 seconds; and
109 72°C 2 minutes.
110
111 **Rolling circle amplification**
112 Each Phi29-dependent rolling circle amplification on dumbbells was primed with
113 loop-hybridizing primer SL or SL10Ta. Specifically, a 10 µL mixture of 10 fmol
114 dumbbell template, 0.5 or 5 pmol primer, and 1mM dNTP was heated at 95°C for 3

115 minutes before ramping down slowly to 25°C. Bovine serum albumin at final 0.2
116 mg/mL and Phi29 polymerase (New England BioLabs) were added and the reaction
117 ran at 30°C for 16 hours before inactivation at 65°C for 10 minutes.
118
119 **Tunneled PacBio sequencing**
120 Customized primers to amplify full-length 16S with P5/P7 ends are illustrated in
121 **Fig. 3a**. Sequence details are in **Supplementary Table 3**. PCR thermocycles were:
122 95°C 3 minutes, 25-28 cycles of 95°C 30 seconds + 57°C 30 seconds (ramp rate ≤ -3
123 °C/sec) + 72°C 1 minute, and 72°C 3 minutes. Each 25 µl reaction contained 2 ng
124 template DNA, 0.375 µM primer each, and KAPA HiFi HotStart ReadyMix (Roche).
125 PCR products were cleaned up with MinElute PCR Purification Kit (QIAGEN) and
126 quantified by DNA Qubit™ dsDNA HS Assay Kit (Invitrogen™). Purified amplicons
127 receiving the same pair of SLOs were pooled equally before assembly. Most
128 samples were subjected to asymmetric looping (**Supplementary Table 1**), which
129 allowed docking of the PacBio sequencing primer on the P5 loop only. One sample
130 received symmetric looping, which took the PacBio sequencing primer on both loops.
131 SLOs for asymmetric looping are Pho5'-C*T*C*T*C*TCAACAACAAC
132 TCCTCCTCCTCCGTT GTTGTTGTTGAGAGAGAT GATAC GGCGACCACCGAGATCTACAC-3'
133 and Pho5'-G*A*A*G*T*GCGCTG TAAGTATTAC CAGCGCACTTC ATAT
134 CAAGCAGAAGACGGCATACGAGAT-3'. The latter was replaced with Pho5'-
135 C*T*C*T*C*TCAACAACAAC TCCTCCTCCTCCGTT GTTGTTGTTGAGAGAGAT CAA
136 GCAGAAGACGGCATACGAGAT -3' for symmetric looping. The assembly condition
137 has been described above in 'Dumbbell formation.' Two samples were from DH5α
138 *E. coli* strain, one sample was ZymoBIOMICS Microbial Community Standard (D6300,
139 ZYMO RESEARCH), and the others were controls that have been sequenced on the
140 PacBio platform before.
141
142 **Bioinformatic analysis of tunneled PacBio sequencing**
143 In brief, adapters were recalled using PacBio's utility *recalladapters*. The output
144 subreads.bam was processed by the *ccs* utility to get the fastq file. After
145 demultiplexing, records belonging to the standard microbial community were pooled
146 into another fastq file for downstream analyses following the published workflow.
147 Detailed commands and scripts are in the supplementary file.
148
149 **Results**
150 **One-step making of DNA dumbbells**
151 This study followed Gibson Assembly[9] in utilizing 5'-exonucleases to expose sticky
152 ssDNA in fixed P5/P7 sequences on Illumina libraries (**Fig. 1a**). Introduction of self-

153 looping oligonucleotides with complementary 3'-ends completed dumbbell
154 formation in one pot. Phosphorothioate bonds were artificially introduced at the
155 5'-ends of looping oligonucleotides to resist the action of 5' exonucleases.
156 As a demonstration of the method, five sequencing libraries of DH5α 16S with
157 P5/P7 ends were prepared, with lengths varying over four powers of two (**Fig. 1b**, left
158 panel). Library conversions were performed using commercially-available ready-mix
159 NEBBuilder kits. Purified products were subjected to digestion by a cocktail of
160 exonucleases (**Fig. 1b**, right panel). Only successfully converted dumbbells were not
161 degraded by the exonuclease cocktail. Conversion efficiency decreased moderately
162 with input length, ranging from 69% for the longest input to 98% for the shortest
163 input. Without conversions, input DNA was highly susceptible to exonucleases;
164 only 1.4% of unconverted input DNA remained undegraded in the longest input (**Fig.
165 1b**, right panel).
166
167 **Verification of dumbbell formation**
168 Looping reactions were verified using PCR (**Fig. 2a**). Dumbbells with known loop
169 sequences were prepared from the DH5α 16S sequence. Two primers, SL and
170 SL10Ta, differ in end sequences but hybridize with the same loop of each dumbbell.
171 Both were each paired with a primer matching the input sequence (16S-V4-R). Only
172 dumbbells, and not unmodified inputs, could yield positive bands. SL and SL10Ta
173 were also used for Phi29-based rolling circle amplification[2] (**Fig. 2b**). Only
174 dumbbells, and not unmodified inputs, allowed Phi29-catalyzed polymerization of
175 high molecular-weight products (**Fig. 2b**, arrow).
176
177 **Tunneling to PacBio Sequencing**
178 DNA dumbbells are the template format of PacBio sequencers. Analyses of a
179 standard microbial community on the PacBio platform have been published by
180 Callahan et al[10]. For comparison, we prepared a full-length 16S library from the
181 same standard community with customized PCR primers (**Fig. 3a and Supplementary
182 Table 1**), along with other full-length 16S control libraries (**Supplementary Table 2**).
183 These PCR oligos differed from PacBio standards on the 5' ends, where the forward
184 primer had Illumina P5 sequences and the reverse primer had Illumina P7 sequences,
185 respectively. After PCR amplification, these linear DNA libraries were converted to
186 dumbbell derivatives.
187 The libraries were sequenced on the PacBio Sequel platform with default
188 parameters (**Supplementary Table 2**). Demultiplexed reads of the standard
189 community were analyzed with the same pipeline from Callahan et al[10]. 23
190 amplicon sequence variants (ASV) of full-length 16S were identified by DADA2[11], 12

191 of which matched exactly to reference sequences (**Supplementary Table 3**). All
192 eight bacteria genera, including *Bacillus, Enterococcus, Escherichia, Lactobacillus,*
193 *Listeria, Pseudomonas, Salmonella, and Staphylococcus* were detected. Their
194 genomic abundances are supposed to be equal. Our results reasonably presented
195 the prediction with a relatively even distribution (**Fig. 3b**). Precise stoichiometry of
196 ASV per genome (**Fig. 3c**) corresponding to Callahan's results[10] was successfully
197 reproduced.
198
199 **Discussion**
200 In this study we offered a simple method to convert P5/P7-ended short-read
201 libraries into DNA dumbbells in one step. Resultant dumbbells are ready for Phi29-
202 based rolling circle amplifications (RCA)[2]. PacBio uses RCA as the *bona fide*
203 sequencing mechanism. Oxford Nanopore can employ RCA to increase sequencing
204 accuracies[12, 13]. MGI Tech needs RCA to prepare sequencing templates as 'DNA
205 nanoballs'[4]. In other words, our method provides a unified one-step option to
206 tunnel all P5/P7-ended sequencing libraries to a long-read format which operates on
207 many different sequencers.
208 Known methods to make DNA dumbbells often take multiple steps and employs
209 PCR[14, 15]. Instead our method combines the principle of Gibson Assembly[9] and the
210 unique topology of self-looping oligonucleotides to make DNA dumbbells in one step.
211 This protocol modifies only the end nucleotides of input libraries. Native
212 nucleotides inside flanking P5/P7 sequences can in principle be left untouched.
213 Asymmetric dumbbells can be readily achieved by ligating P5- or P7-specific loops.
214 Methods for linking multiple dumbbells in tandem can be pursued with minor
215 modifications.
216 We demonstrated the tunneling option to the PacBio sequencer (**Fig. 3**).
217 Although read depths were suboptimal (**Supplementary Table 2**), analyses of the
218 standard microbial community exhibited very good correspondences to previously
219 published results[10]. We saw reasonably distributed abundances of bacteria
220 genomes from the mock standard community. Totally 23 ASV of full-length 16S
221 were identified. The number was six fewer than that identified by Callahan's[10] et
222 all. With a significant difference in sequencing depths, the disparity was not
223 unexpected. Importantly, stoichiometry of ASV per genome matched Callahan's
224 results exactly. The missing three ASV for *Bacillus*, two ASV for *Enterococcus*, and
225 one ASV for *Staphylococcus* did not change overall integral sums of ASV for these
226 bacteria, which remained 10, 4, and 6, respectively.
227
228 **Acknowledgements**

We thank Chih-Wei Joshua Liu (MIT Physics of Living Systems) for English editing. Human samples previously sequenced on PacBio sequencers were kindly provided by Dr. Yu-Chen Lin at Far-Eastern Memorial Hospital. This work was supported by National Health Research Institutes (MG-109-SP-05).

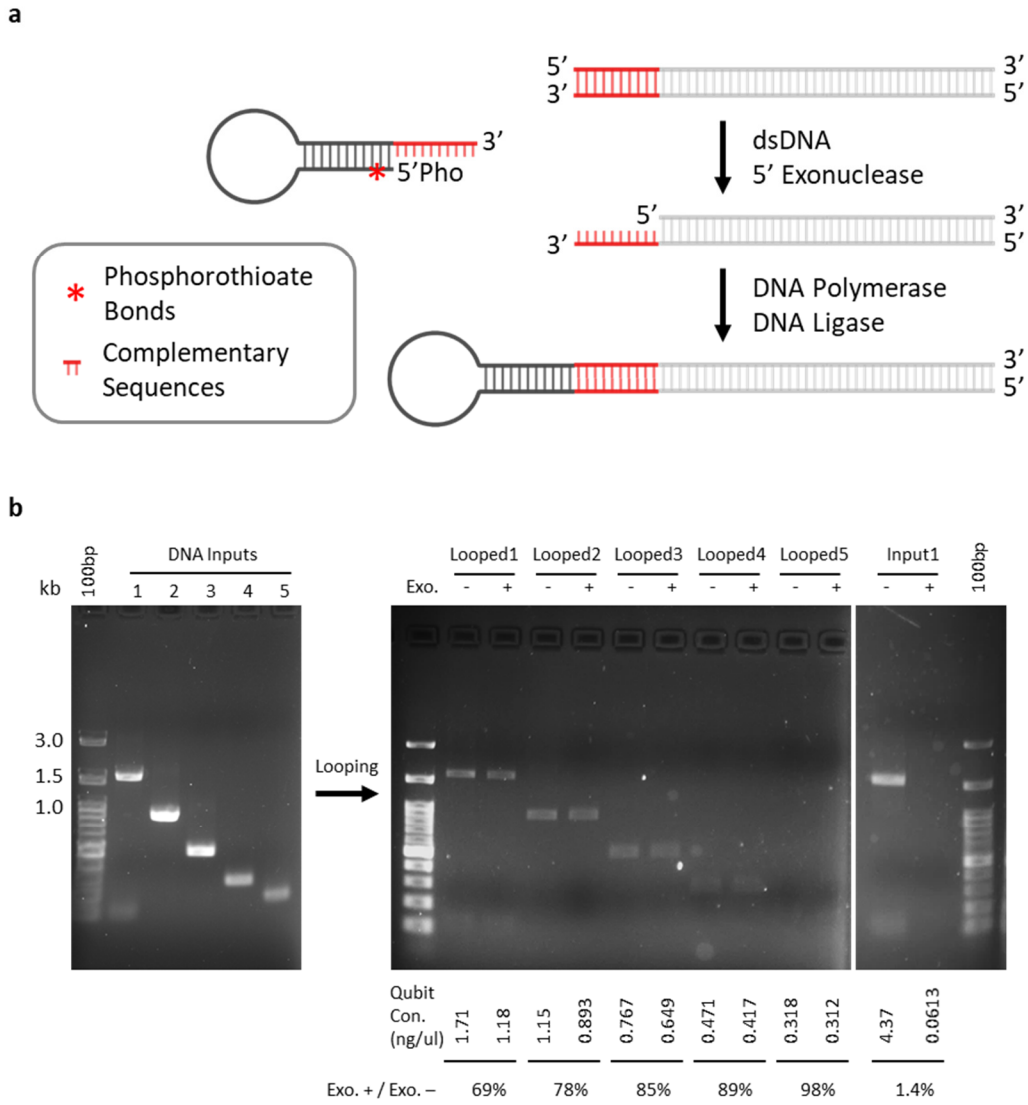
Author Contributions

Y.H.C. conducted experiments. H.H.L. invented the method.

Competing Financial Interests

A patent application is being filed.

240 **Figures**



241
242

243 **Figure 1** | Demonstration of protocol. (**a**) P5/P7 ends of input libraries are processed
244 using a 5'-exonuclease, as in Gibson Assembly. Only one end is illustrated. 3'-ssDNA is
245 exposed by dsDNA 5'-exonuclease, which allows 5'-phosphorothioated and 5'-
246 phosphorylated self-looping oligonucleotides to anneal with complementary 3'-
247 overhangs. DNA polymerase and ligase complete the looping reaction. (**b**) Five DH5α
248 16S libraries of lengths varying by factors of two were subjected to conversion.
249 Products were tested with a cocktail of exonucleases, which has no effect on
250 successfully looped dumbbells. Conversion efficiencies varied from 69% for the
251 longest input sequence to 98% for the shortest input sequence. Little unconverted
252 input survived digestion by the exonuclease cocktail.
253

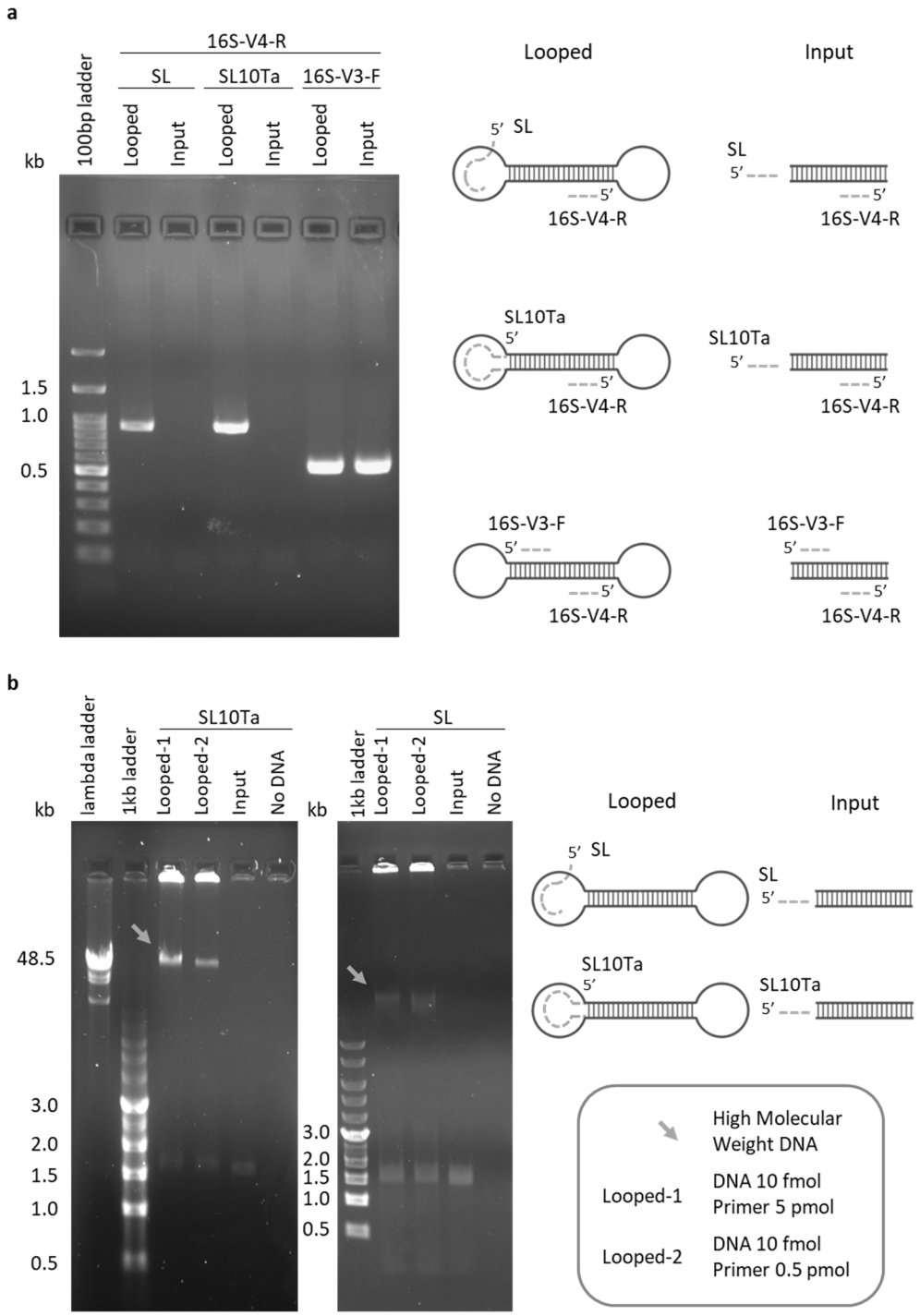


Figure 2 | Verification of protocol. (**a**) Two different primers capable of hybridizing with a loop paired with a reverse primer matching the input sequence. Only looped products, and not unconverted inputs, yielded positive PCR bands. (**b**) Rolling circle amplifications by Phi29 polymerase proceeded only on dumbbell templates primed with two different loop-hybridizing primers. No high molecular-weight products (arrow) were produced using unconverted inputs as templates.

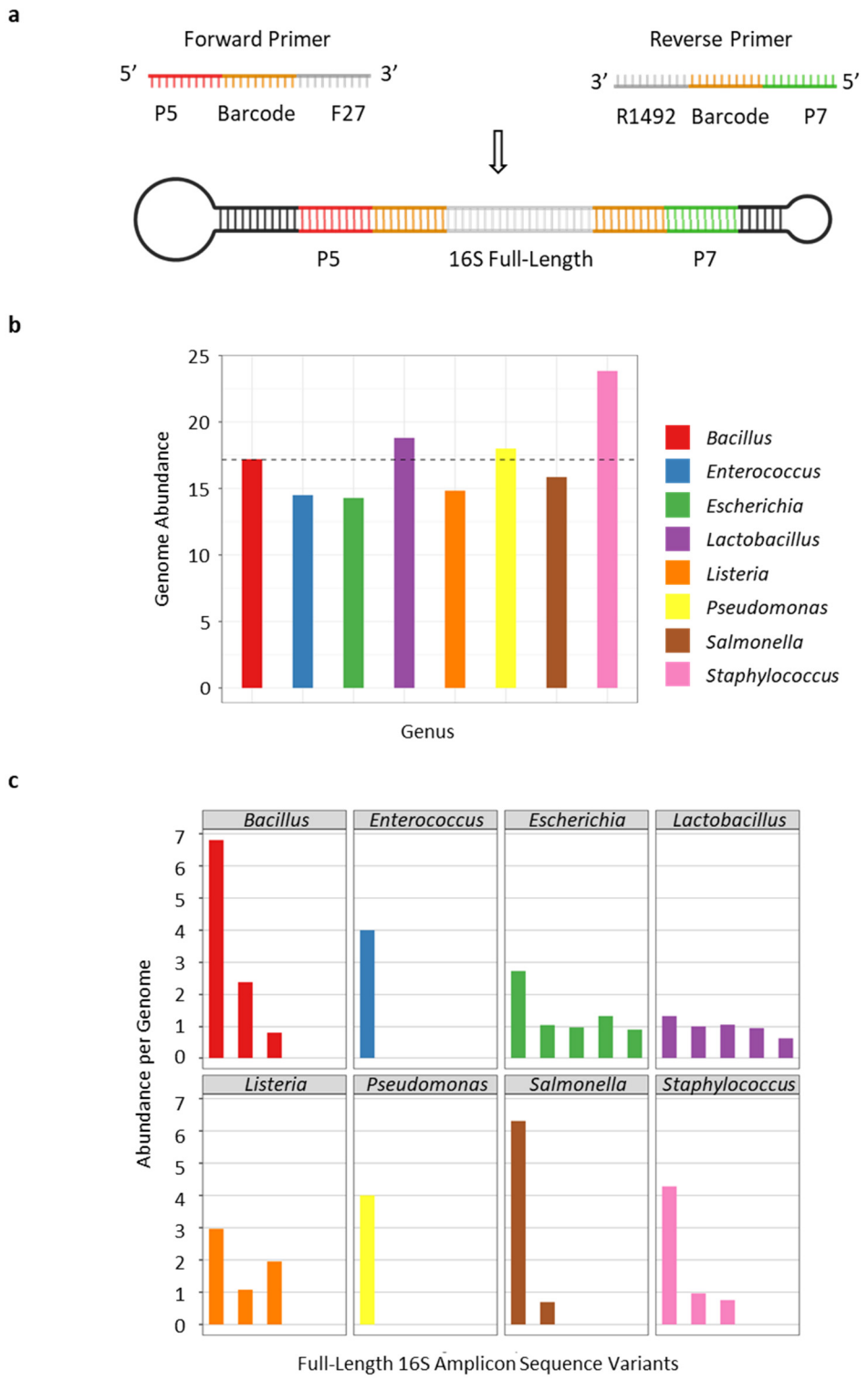


Figure 3 | Tunneling to PacBio sequencing. (**a**) A pair of primers able to amplify full-length 16S with P5/P7 end sequences are illustrated. (**b**) Genome abundances of eight bacteria genera were relatively equal, corresponding to the made composition of the microbial community standard. (**c**) Stoichiometric analyses of 16S ASV per bacteria genome revealed integral values and sums in line with expectations.

269 **References**

270 1. Hecker, M. & Wagner, A.H. Transcription factor decoy technology: A
271 therapeutic update. *Biochem Pharmacol* **144**, 29-34 (2017).
272 2. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of
273 plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed
274 rolling circle amplification. *Genome Res* **11**, 1095-1099 (2001).
275 3. Hon, T. et al. Highly accurate long-read HiFi sequencing data for five complex
276 genomes. *Sci Data* **7**, 399 (2020).
277 4. Drmanac, R. et al. Human genome sequencing using unchained base reads on
278 self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).
279 5. Gorzynski, J.E. et al. Ultrarapid Nanopore Genome Sequencing in a Critical
280 Care Setting. *N Engl J Med* **386**, 700-702 (2022).
281 6. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly
282 multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*
283 **2010**, pdb. prot5448 (2010).
284 7. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. A flexible and
285 efficient template format for circular consensus sequencing and SNP
286 detection. *Nucleic Acids Res* **38**, e159 (2010).
287 8. Slatko, B.E., Gardner, A.F. & Ausubel, F.M. Overview of Next-Generation
288 Sequencing Technologies. *Curr Protoc Mol Biol* **122**, e59 (2018).
289 9. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several
290 hundred kilobases. *Nat Methods* **6**, 343-345 (2009).
291 10. Callahan, B.J. et al. High-throughput amplicon sequencing of the full-length
292 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res* **47**, e103
293 (2019).
294 11. Callahan, B.J. et al. DADA2: High-resolution sample inference from Illumina
295 amplicon data. *Nat Methods* **13**, 581-583 (2016).
296 12. Volden, R. et al. Improving nanopore read accuracy with the R2C2 method
297 enables the sequencing of highly multiplexed full-length single-cell cDNA.
298 *Proc Natl Acad Sci U S A* **115**, 9726-9731 (2018).
299 13. Wilson, B.D., Eisenstein, M. & Soh, H.T. High-Fidelity Nanopore Sequencing of
300 Ultra-Short DNA Targets. *Anal Chem* **91**, 6783-6789 (2019).
301 14. Taki, M. et al. A direct and efficient synthesis method for dumbell-shaped
302 linear DNA using PCR in vitro. *Nucleic Acids Res Suppl*, 191-192 (2003).
303 15. Yu, H., Jiang, X., Tan, K.T., Hang, L. & Patzel, V. Efficient production of superior
304 dumbbell-shaped DNA minimal vectors for small hairpin RNA expression.
305 *Nucleic Acids Res* **43**, e120 (2015).

306