A One-Step Synthesis of DNA Dumbbells Tunnels 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. For example, they can be templates for rolling-circle amplification to produce highly accurate long-read sequences of 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. Resulting dumbbells were ready for Phi29-dependent rolling circle amplification. Demonstrative readouts of full-length 16S rDNA from a standard microbial community, obtained on a Pacific Biosciences (PacBio) platform, confirmed successful tunneling of initially P5/P7-ended sequencing libraries. Twelve fecal samples additionally showed significant correlations between standard and tunneled 16S sequence variants on a PacBio platform.

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

Introduction

DNA dumbbells are a variety of circular DNA. Topologically equivalent to circles, DNA dumbbells are highly resistant to exonucleases and can serve as stable vectors for regulating gene expression[1]. Additionally, dumbbell DNA can serve as an iterative template for rolling-circle amplification[2]. Many modern sequencers use DNA dumbbells or circles as sequencing templates. Pacific Biosciences (PacBio) platforms, for example, derive long and highly accurate sequences *via* alignment of iterative 'subreads' into consensus 'HiFi' reads[3]. MGI Tech instead uses circular DNA to amplify sequencing signals *via* 'DNA nanoballs'[4].

With high-throughput sequencing, a complete human genome can now be analyzed within hours[5]. Decreased costs and increased throughputs have accelerated

application of sequencing technology to 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 advantages and disadvantages[8]. Illumina platforms are highly accurate but suffer from short reads of 150bp paired-end on most recent hardware. PacBio platforms instead offer long reads of well over 10kbp, achieving accuracy through iteratively optimized 'HiFi' consensus[3]. For some libraries, such as those that include full-length B-lymphocyte heavy chain sequences, neither platform is readily used. These libraries are typically longer than what Illumina platforms can comfortably sequence, but few established workflows exist to format these libraries for PacBio platforms. A potential solution unifies the protocols through a 'tunneling' option. We have realized this solution with a novel one-step synthesis 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 10 µL pellet from a 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 NEBuilder® HiFi DNA Assembly Master Mix (New England BioLabs®) at 45°C for 2-3 hours. The 5'-ends of SLOs were phosphorylated and protected by five consecutive

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

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

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

191 Their genomic abundances are expected to be equal. Our results recapitulated the
192 prediction with a relatively even distribution (**Fig. 3b**). Precise stoichiometry of ASV
193 per genome (**Fig. 3c**) corresponding to Callahan's results[10] was successfully
194 reproduced.
195 Twelve fecal full-length 16S libraries previously sequenced following PacBio's
196 standard protocol (**Supplementary Table 2**), henceforth referred to as direct
197 sequencing (DS), were re-sequenced as controls. ASV of each sample by DS and
198 tunneled sequencing (TS) were both identified by DADA2[11]. Although DS had higher
199 HiFi read depths than TS (14.4 fold on average), ASV identified from DS were only 4.5
200 fold higher than TS (**Table 2**). Among TS ASV, 92.2±2.1% (s.e.) overlapped with those
201 from DS. Abundances of overlapped ASV showed significant correlations (**Fig. 3d**).
202 Pearson coefficients ranged from 0.77 to 0.99 (0.89±0.02, s.e.), with a significance
203 level $<10^{-6}$ at least.
204
205 **Discussion**
206 In this study, we demonstrate a simple method to convert P5/P7-ended short-
207 read libraries into DNA dumbbells in one step. Resulting dumbbells are ready for
208 Phi29-based rolling circle amplification (RCA)[2]. PacBio uses RCA as its *bona fide*
209 sequencing mechanism. Oxford Nanopore can employ RCA to increase sequencing
210 accuracy[12, 13]. MGI Tech uses RCA to prepare sequencing templates as 'DNA
211 nanoballs'[4]. In other words, our method provides a unified one-step option for
212 tunneling any P5/P7-ended sequencing library to a long-read format suitable for
213 many different sequencers.
214 Previous methods for forming DNA dumbbells use multiple steps and employ
215 PCR[14, 15]. Our method instead combines the principle of Gibson Assembly[9] and the
216 unique topology of self-looping oligonucleotides to make DNA dumbbells in one step.
217 This protocol modifies only the end nucleotides of input libraries. Native nucleotides
218 inside flanking P5/P7 sequences can in principle be left untouched. Asymmetric
219 dumbbells can be readily achieved by ligating P5- or P7-specific loops. Methods for
220 linking multiple dumbbells in tandem can be pursued with minor modifications.
221 We demonstrated the tunneling option on a PacBio sequencer (**Fig. 3**). Although
222 read depths were suboptimal (**Table 2**), analyses of the standard microbial
223 community exhibited very good correspondence to previously published results[10].
224 We saw reasonably distributed abundances of bacterial genomes from the mock
225 standard community. In total, 23 full-length 16S ASV were identified. The number
226 was six fewer than identified by Callahan et al[10]. Due to the significant difference in
227 sequencing depths, this disparity is expected. Importantly, stoichiometry of ASV per
228 genome matched Callahan's results exactly. The missing three ASV for *Bacillus*, two

ASV for *Enterococcus*, and one ASV for *Staphylococcus* did not change overall integral sums of ASV for these genera, which remained 10, 4, and 6, respectively. Even for fecal samples with complex bacterial communities, reads of tunneled libraries correlated well (**Fig. 3d**).

The reason for suboptimal read depths from tunneled sequencing on the PacBio platform is unknown. Potential explanations include sequencing primers mismatching our customized loop, unoptimized stoichiometry of sequencing primers *vs.* asymmetric dumbbell templates, and inefficient docking of the primer-dumbbell complex. All of these can be corrected with fine-tuning of our technique.

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. developed the method.

Competing Financial Interests

A patent application is being filed.

251 **Table 1** Matches of ASV to ZymoBIOMICS reference

Genus	No	Yes
Bacillus	0	3
Enterococcus	0	1
Escherichia	5	0
Lactobacillus	5	0
Listeria	0	3
Pseudomonas	0	1
Salmonella	0	2
Staphylococcus	1	2

252

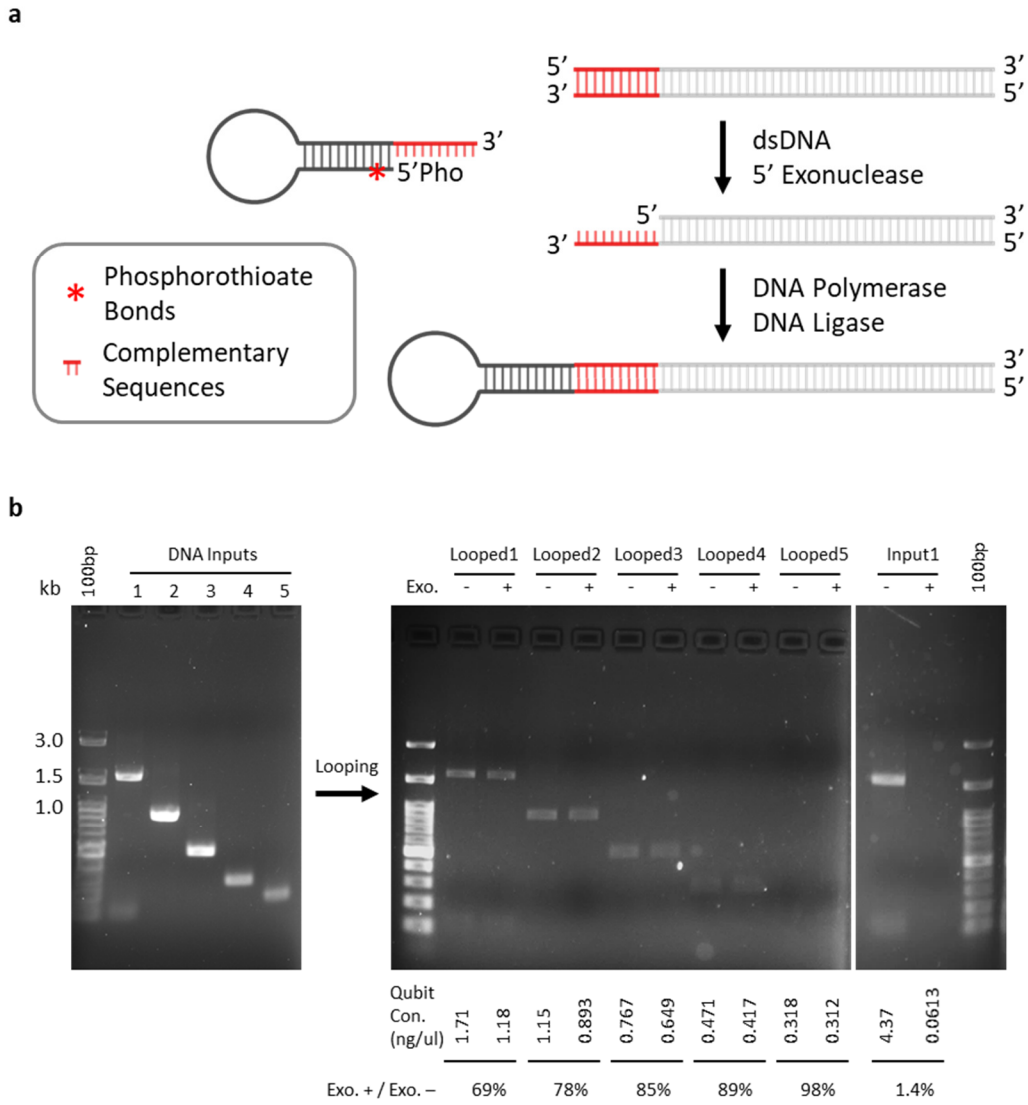
253

254

Table 2 Summary of HiFi reads and ASV from direct and tunneled sequencing

Sample	HiFi Reads by Direct Sequencing	ASV by Direct Sequencing	HiFi Reads by Tunneled Sequencing	ASV by Tunneled Sequencing	Matched ASV between Both Sequencings
#04	11009	184	864	44	43
#05	10217	197	881	55	46
#06	29282	295	889	38	33
#07	11429	167	858	34	32
#09	10193	318	866	32	30
#10	29334	133	856	38	29
#11	10146	431	822	53	51
#12	30917	159	1183	40	40
#13	12164	95	703	31	28
#14	11543	167	747	31	28
#15	11730	144	633	29	28
#67	11694	55	1180	17	17

257 **Figures**



258
259

260 **Figure 1** | Demonstration of protocol. (**a**) P5/P7 ends of input libraries are processed
261 using a 5'-exonuclease, as in Gibson Assembly. Only one end is illustrated. 3'-ssDNA is
262 exposed by dsDNA 5'-exonuclease, which allows 5'-phosphorothioated and 5'-
263 phosphorylated self-looping oligonucleotides to anneal with complementary 3'-
264 overhangs. DNA polymerase and ligase complete the looping reaction. (**b**) Five DH5α
265 16S libraries of lengths varying by factors of two were subjected to conversion.
266 Products were tested with a cocktail of exonucleases, which has no effect on
267 successfully looped dumbbells. Conversion efficiencies varied from 69% for the
268 longest input sequence to 98% for the shortest input sequence. Little unconverted
269 input survived digestion by the exonuclease cocktail.
270

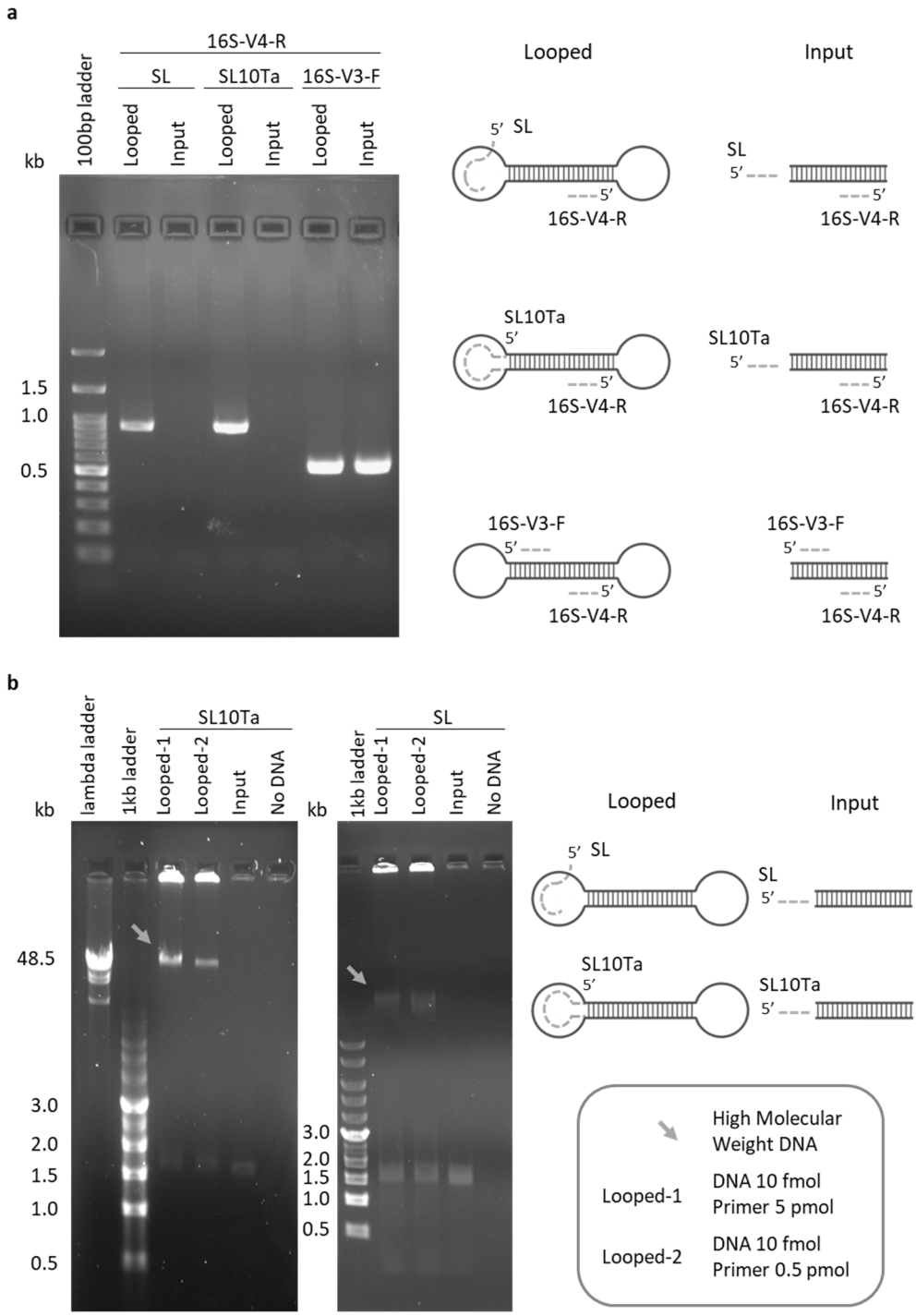


Figure 2 | Verification of protocol. (**a**) Two different primers capable of hybridizing with a loop paired with a reverse primer matching the input sequence. Only looped products, and not unconverted inputs, yielded positive PCR bands. (**b**) Rolling circle amplifications by Phi29 polymerase proceeded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

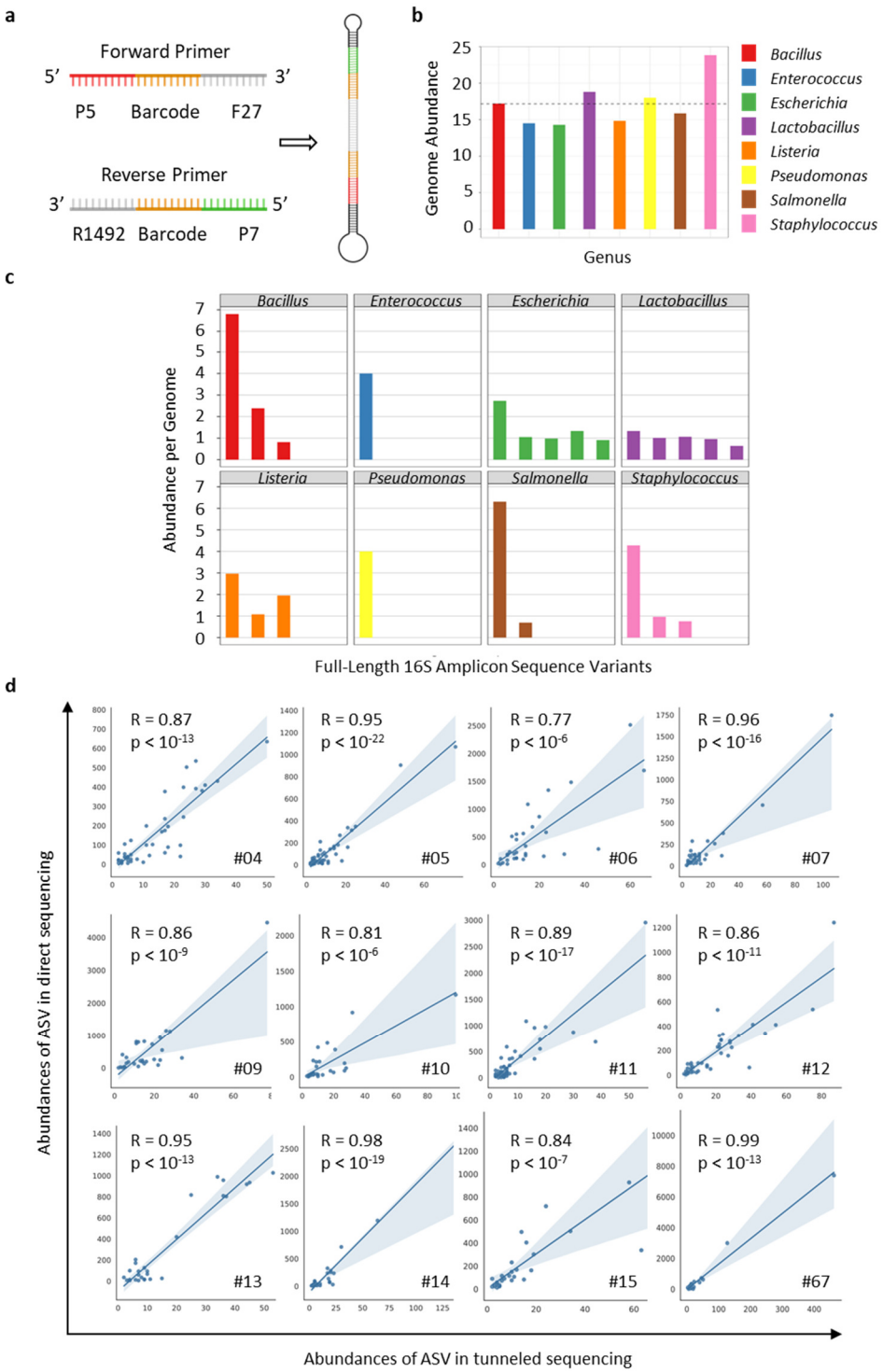


Figure 3 | Tunneling to PacBio sequencing. (**a**) A pair of primers able to amplify full-length 16S with P5/P7 end sequences is illustrated. (**b**) Genome abundances of eight bacterial genera were relatively equal, corresponding to the known composition of the microbial community standard. (**c**)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

288 **References**

289 1. Hecker, M. & Wagner, A.H. Transcription factor decoy technology: A
290 therapeutic update. *Biochem Pharmacol* **144**, 29-34 (2017).
291 2. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of
292 plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed
293 rolling circle amplification. *Genome Res* **11**, 1095-1099 (2001).
294 3. Hon, T. et al. Highly accurate long-read HiFi sequencing data for five complex
295 genomes. *Sci Data* **7**, 399 (2020).
296 4. Drmanac, R. et al. Human genome sequencing using unchained base reads on
297 self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).
298 5. Gorzynski, J.E. et al. Ultrarapid Nanopore Genome Sequencing in a Critical
299 Care Setting. *N Engl J Med* **386**, 700-702 (2022).
300 6. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly
301 multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*
302 **2010**, pdb. prot5448 (2010).
303 7. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. A flexible and
304 efficient template format for circular consensus sequencing and SNP
305 detection. *Nucleic Acids Res* **38**, e159 (2010).
306 8. Slatko, B.E., Gardner, A.F. & Ausubel, F.M. Overview of Next-Generation
307 Sequencing Technologies. *Curr Protoc Mol Biol* **122**, e59 (2018).
308 9. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several
309 hundred kilobases. *Nat Methods* **6**, 343-345 (2009).
310 10. Callahan, B.J. et al. High-throughput amplicon sequencing of the full-length
311 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res* **47**, e103
312 (2019).
313 11. Callahan, B.J. et al. DADA2: High-resolution sample inference from Illumina
314 amplicon data. *Nat Methods* **13**, 581-583 (2016).
315 12. Volden, R. et al. Improving nanopore read accuracy with the R2C2 method
316 enables the sequencing of highly multiplexed full-length single-cell cDNA.
317 *Proc Natl Acad Sci U S A* **115**, 9726-9731 (2018).
318 13. Wilson, B.D., Eisenstein, M. & Soh, H.T. High-Fidelity Nanopore Sequencing of
319 Ultra-Short DNA Targets. *Anal Chem* **91**, 6783-6789 (2019).
320 14. Taki, M. et al. A direct and efficient synthesis method for dumbell-shaped
321 linear DNA using PCR in vitro. *Nucleic Acids Res Suppl*, 191-192 (2003).
322 15. Yu, H., Jiang, X., Tan, K.T., Hang, L. & Patzel, V. Efficient production of superior
323 dumbbell-shaped DNA minimal vectors for small hairpin RNA expression.
324 *Nucleic Acids Res* **43**, e120 (2015).

325