Measuring Illumina Size Bias Using REcount: A Novel Method for Highly Accurate Quantification of Engineered Genetic Constructs

Daryl M. Gohl[1,2,4], Aaron Becker[1], Darrell Johnson[1], Shea Anderson[1], Bradley Billstein[1], Shana L. McDevitt[3], Kenneth B. Beckman[1]

[1]University of Minnesota Genomics Center, Minneapolis, MN, 55455, USA
[2]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA
[3]Vincent J. Coates Genomics Sequencing Laboratory, University of California, Berkeley, CA, 94720, USA
[4]Corresponding author: dmgohl@umn.edu

Keywords: Next-generation sequencing, DNA library preparation, PCR-free, Illumina, size bias, clustering.

Abstract

Quantification of DNA sequence tags associated with engineered genetic constructs underlies many genomics measurements. Typically, such measurements are done using PCR to enrich sequence tags and add adapters, followed by next-generation sequencing (NGS). However, PCR amplification can introduce significant quantitative error into these measurements. Here we describe REcount, a novel PCR-free direct counting method for NGS-based quantification of engineered genetic constructs. By comparing measurements of defined plasmid pools to droplet digital PCR data, we demonstrate that this method is highly accurate and reproducible. We further demonstrate that the REcount approach is amenable to multiplexing through the use of orthogonal restriction enzymes. Finally, we use REcount to provide new insights into clustering biases due to molecule length across different Illumina sequencing platforms.

Introduction

Engineered genetic constructs underlie many experimental techniques in genetics and genomics. For example, targeted perturbation of gene function using RNA interference or CRISPR/Cas9 allows for pooled genome-wide genetic screens that can be read-out through next-generation sequencing (NGS) of the small hairpin RNA (shRNA) (Sims et al. 2011; Rodriguez-Barrueco et al. 2013) or synthetic guide RNA (sgRNA) (Wang et al. 2014; Koike-Yusa et al. 2014; Shalem et al. 2014, 2015) constructs, or associated sequence tags/barcodes (Smith et al. 2009). Transposable elements are also commonly used to mutate or otherwise manipulate genetic loci, and similarly enable genome-scale saturation mutagenesis screens in which the transposon-genome junction is measured using NGS (van Opijnen and Camilli 2013). Lineage-tracing (Bhang et al. 2015; McKenna et al. 2016) and connectomics (Kebschull et al. 2016; Peikon et al. 2017) approaches also rely on NGS-based quantification of molecular tags. In all of these approaches, Polymerase Chain Reaction (PCR) amplification is used to enrich for the sequence tags and to add adapters and other functionalities (e.g. sample-specific barcodes) required for sequencing. However, PCR introduces bias into these measurements. Sequence tags comprised of shRNAs, sgRNAs, transposon:genome junctions, or synthetic barcodes, can all differ in primary sequence and biophysical properties, which, along with other variables such as template concentration and PCR conditions, can influence amplification efficiency in unpredictable ways (Aird et al. 2011; Gohl et al. 2016; Strezoska et al. 2012). Adding unique molecular identifiers (UMIs) can mitigate some of this bias, but increases the complexity of both library preparation and analysis (Kivioja et al. 2011; Kinde et al. 2011). Other approaches such as droplet digital PCR (ddPCR) and NanoString analysis can be used to overcome the quantitative inaccuracies associated with measuring engineered genetic constructs, but lack the throughput and resolution afforded by NGS (Geiss et al. 2008; Hindson et al. 2011).

We have developed a novel method, REcount (**R**estriction **E**nzyme enabled **count**ing) for quantifying sequence tags associated with engineered genetic constructs that is straightforward to implement and allows for direct NGS-based counting of a potentially enormous number of sequence tags. In this approach, an Illumina adapter-flanked DNA barcode is liberated by digesting with *Mly*I (a type IIS restriction enzyme that produces blunt-ended molecules) and sequenced to directly count template molecule abundance (Figure 1A). We use REcount to design a set of synthetic DNA standards that can be used to assess clustering bias due to molecule length on Illumina sequencers, and demonstrate that there is substantial variation in size bias between different Illumina instruments.

Results

In order to characterize the REcount method, we constructed a pool of 20 synthetic plasmids containing REcount barcodes, mixed at an equimolar abundance (5% per plasmid) based on fluorometric DNA concentration measurements. This pool was digested with *Mly*I and sequenced on an Illumina MiSeq. All 20 barcodes were detected at relative abundances ranging from 3.41% and 6.32% (CV = 0.13), consistent with the targeted abundances of 5% per

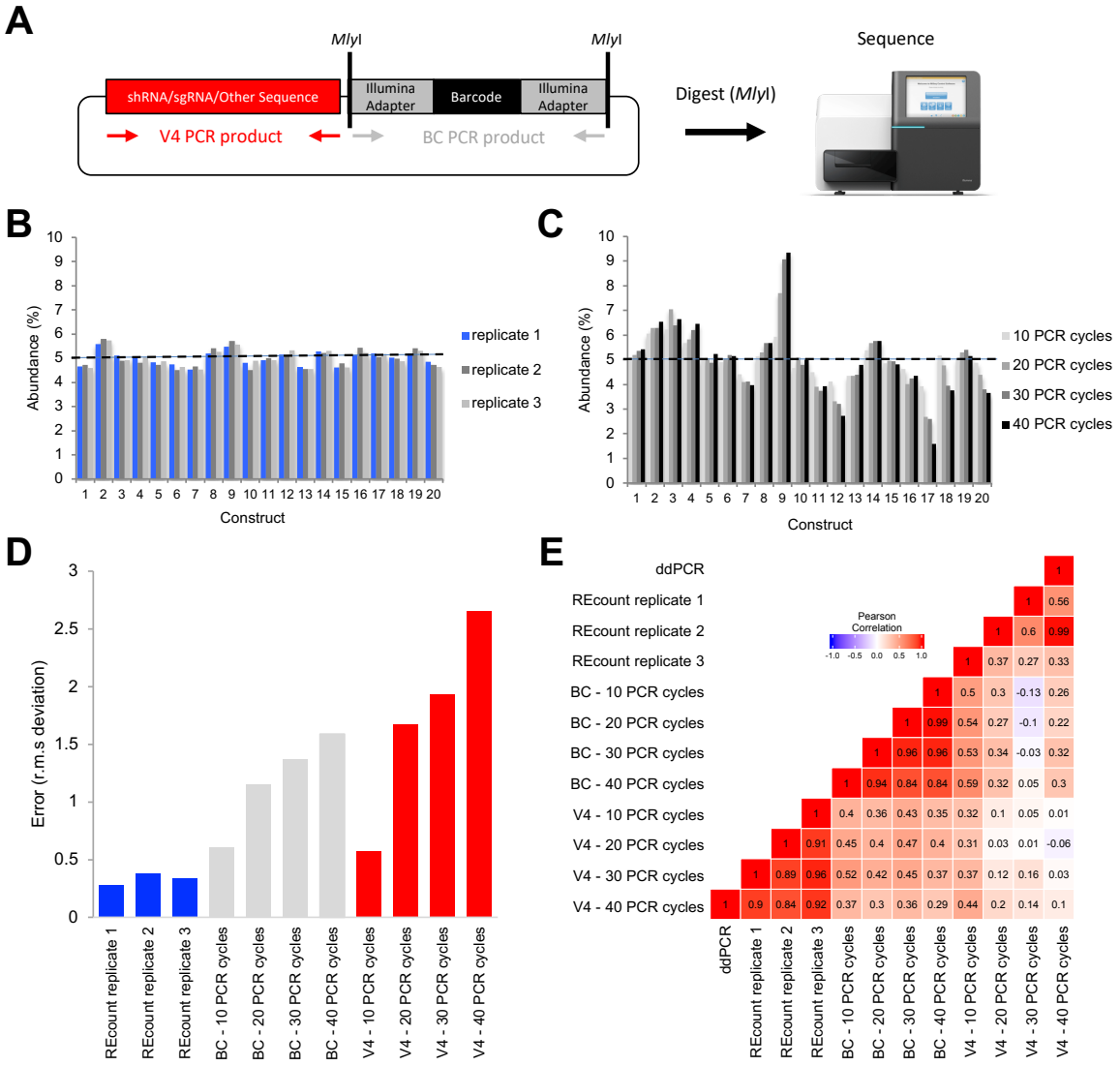

Figure 1. REcount enables accurate and precise measurements of plasmid pools. A) Design of REcount constructs. A barcode-containing, Illumina adapter-flanked construct is liberated with a restriction enzyme (*Mly*I) digest and directly sequenced. B) Accuracy and reproducibility of REcount. C) Analogous measurements of the same plasmid pool shown in panel B using varying PCR cycle numbers. D) Root mean squared deviation from expected values (5% per construct) when the plasmid pool is measured using REcount, and varying cycles of PCR amplification of either the barcode construct (BC) or another variable sequence in these plasmids (V4). E) Pearson correlation heatmap comparing REcount measurements with droplet digital PCR data and with conventional PCR amplification of either the BC or V4 amplicons.

53 construct (Supplemental Figure S1). To generate a more accurately pooled reference standard
54 for subsequent experiments, we used this sequencing data as the basis for re-pooling the 20
55 plasmids and digested the new pool with *Mly*I and sequenced. The range of relative
56 abundances of the re-pooled plasmids was narrower, ranging from 4.52% to 5.58% (CV = 0.06),
57 indicating that the initial sequencing data was predictive in improving the accuracy of pooling as
58 assessed by REcount (Supplemental Figure S1). To assess the reproducibility of these
59 measurements, we digested and sequenced two additional replicates of the even plasmid pool.
60 The replicate REcount measurements were highly reproducible with an average CV of 0.02
61 (Figure 1B).
62 Next, we compared REcount measurements of the even plasmid pool to PCR-based
63 measurements, either of the barcode construct (BC) or another construct-specific sequence
64 (V4). PCR-based measurements exhibited substantial construct-specific deviations from the
65 expected 5% values, the extent of which increased with greater numbers of PCR cycles (Figure
66 1C-D). Furthermore, the construct-specific deviations from expected values were uncorrelated
67 for the BC and V4 amplicon measurements, suggesting that the PCR biases were a function of
68 template sequence (Supplemental Figure S2).
69 In order to independently measure the relative template concentrations in the even
70 plasmid pool, we designed a pair of ddPCR assays targeting each barcode construct and
71 validated the specificity of each assay using qPCR on each of the 20 individual plasmid
72 templates (Supplemental Figure S3) (Hindson et al. 2011). The ddPCR-based measurements
73 correlated well with the REcount measurements, both for the original and re-pooled even
74 plasmid pools (Figure 1E, Supplemental Figure S3). In contrast, the PCR-based measurements
75 of both the BC and V4 amplicons were not well-correlated with the ddPCR measurements
76 (Figure 1E, Supplemental Figure S3). These results were corroborated with similar
77 measurements of a pool of the same 20 plasmids mixed in a staggered manner, where PCR-
78 based measurements had reduced correlation with ddPCR measurements and led to a
79 systematic overestimation of the lower abundance constructs (Supplemental Figure S4). Taken
80 together, these results indicate that REcount accurately reports on template abundance, while
81 PCR-based measurements introduce increasing error with increased cycle numbers.
82 One drawback of the REcount method is that the indices that specify sample identity in
83 multiplexed sequencing, which are typically flexibly added by PCR, are hard-coded into the
84 constructs. To overcome this limitation, we tested whether orthogonal restriction enzymes could
85 be used to multiplex REcount measurements. We initially chose *Mly*I as the flanking enzyme
86 because it could precisely liberate the desired Illumina adapter-flanked construct. We tested

87 whether other restriction enzymes that do not cleanly liberate flush Illumina adapter ends could
88 also be used for REcount measurements. Initially, we tested *Bsm*I, *Bts*$^{\alpha}$I, and *Bsr*DI, each of
89 which leave 2 nt 3' overhangs. We constructed a pool of 12 plasmids comprised of sets of three
90 barcoded constructs flanked by either *Mly*I, *Bsm*I, *Bts*$^{\alpha}$I, or *Bsr*DI (Figure 2A). In addition, all 12
91 of these constructs contained a pair of *Sbf*I sites located such that digestion with *Sbf*I liberates
92 all 12 Illumina adapter-flanked cassettes with additional overhangs of between 30-36 bp
93 upstream of the p5 flowcell adapter and between 40-50 bp downstream of the p7 flowcell
94 adapter. We digested this plasmid pool with each of the 5 enzymes individually and individually
95 sequenced the digests and mapped the reads to a reference file containing all 12 expected

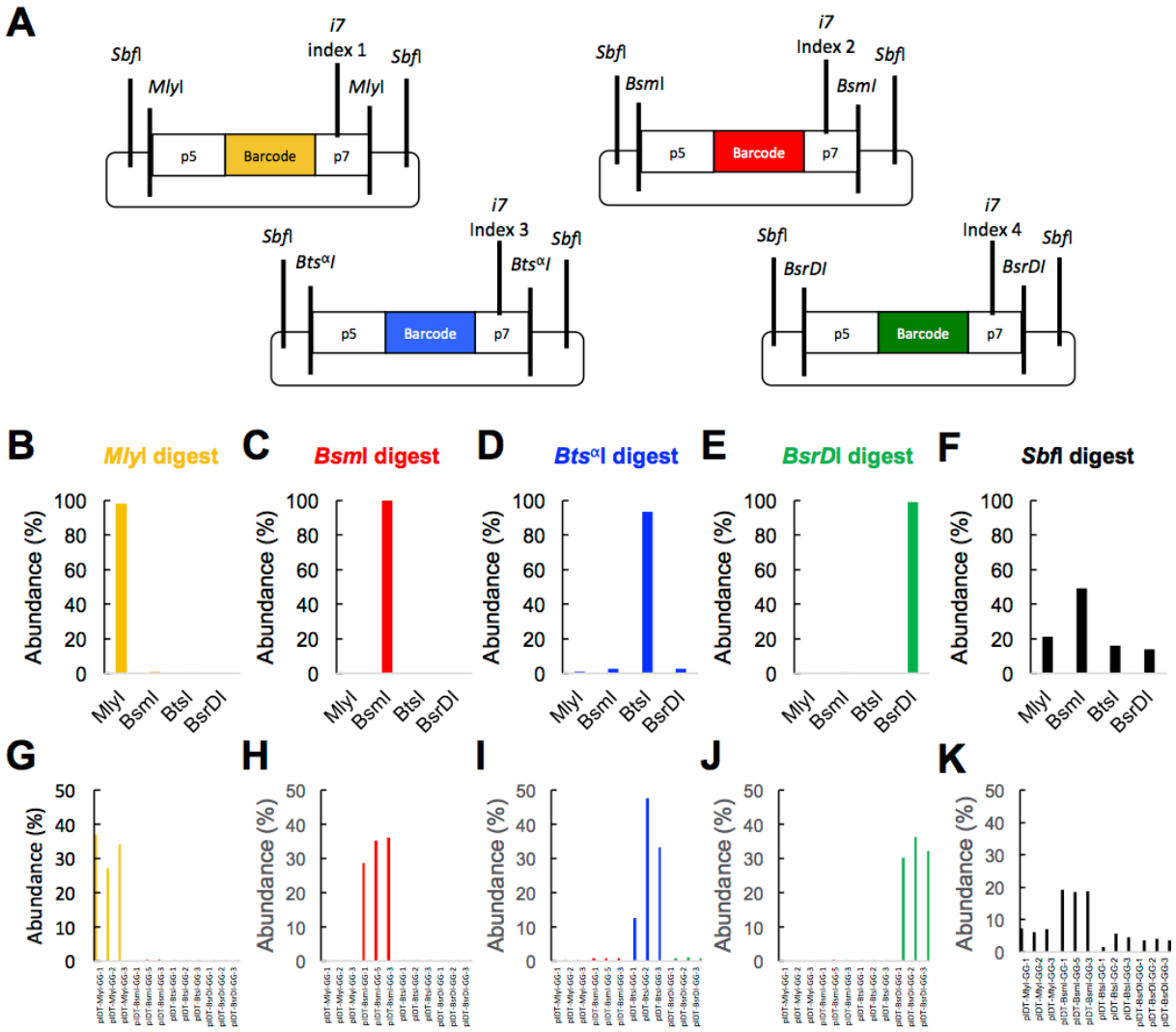

Figure 2. Multiplexing of REcount measurements using orthogonal restriction enzymes. A) Plasmids containing REcount constructs flanked by orthogonal restriction enzyme cut sites. B-F) Total mapped reads identified for each construct type when the plasmid pool is digested with the indicated enzyme. G-K) Mapped reads identified for each construct when the plasmid pool is digested with the indicated enzyme.

96 barcodes. For *Mly*I, *Bsm*I, *Bts*$^{\alpha}$I, and *Bsr*DI, the expected barcodes were detected for each
97 respective enzyme (Figure 2B-E, G-J). All 12 barcodes were detected when the pool was
98 digested with *Sbf*I, indicating that clustering and sequencing can occur even in the presence of
99 large (30-50 bp) overhangs (Figure F, K). We were not able to determine whether the length of
100 the overhang affects the efficiency of clustering as each of these samples was sequenced in a
101 portion of a MiSeq lane, together with other libraries. We observed differing amounts of off-
102 target barcode detection in these orthogonal digests, ranging from <0.2% in the *Bsm*I digest to
103 approximately 6% in the *Bts*$^{\alpha}$I digest (Figure 2B-E, G-J). This could likely be improved by adding
104 a size selection step. Nonetheless, the fact that multiple restriction enzymes can be used to
105 liberate REcount constructs allows for potential multiplexing strategies involving orthogonal
106 digestion of distinct subpopulations of molecules or of concatamerized barcode arrays.
107 While it is known that molecule size affects clustering and sequencing efficiency on
108 Illumina sequencers (Illumina 2014), the extent of this bias and the degree to which it differs
109 between different Illumina instruments has not been characterized in detail. Thus, we used
110 REcount to characterize the size bias profiles of the Illumina MiSeq, HiSeq 2500, HiSeq 4000,
111 NextSeq, and NovaSeq sequencers. We synthesized 30 constructs, each of which contained an
112 *Mly*I-flanked normalization barcode of consistent length (164 bp), and a barcode-containing
113 variable-length insert ranging from 22 bp to 1372 bp, resulting in adapter-flanked molecules
114 between 150 and 1500 bp (Figure 3A). In order to minimize sequence-specific artifacts, the
115 variable-length inserts were chosen to have between 42% and 58% GC content, and were
116 comprised of 10 constructs each (spanning the full 150 bp - 1500 bp size range) derived from
117 three different molecules; the *Escherichia coli* (*E. coli*) 16S rRNA gene (16S), the *Drosophila*
118 *melanogaster* (*D. melanogaster*) *alpha-Tubulin84B* gene (Tubulin), and the *D. melanogaster*
119 *Glyceraldehyde-3-phosphate dehydrogenase 1* (GAPDH) gene (Supplemental Figure S5).
120 These Illumina size standard constructs were pooled at an equimolar ratio based on
121 fluorometric DNA concentration measurements, digested with *Mly*I, and sequenced on different
122 Illumina DNA sequencers with no intervening clean-up step, to ensure that no material was lost.
123 Representative data from a single MiSeq run is shown in Figure 3B. Since each normalization
124 barcode is present at an equimolar ratio to the corresponding size standard (as they are on the
125 same plasmid), this allows any inaccuracies in plasmid pooling to be accounted for. Within a
126 sequencing platform, clustering size bias exhibits run-to-run variation (Figure 3C). All five of the
127 sequencers we tested exhibited preferential clustering of smaller fragments, consistent with

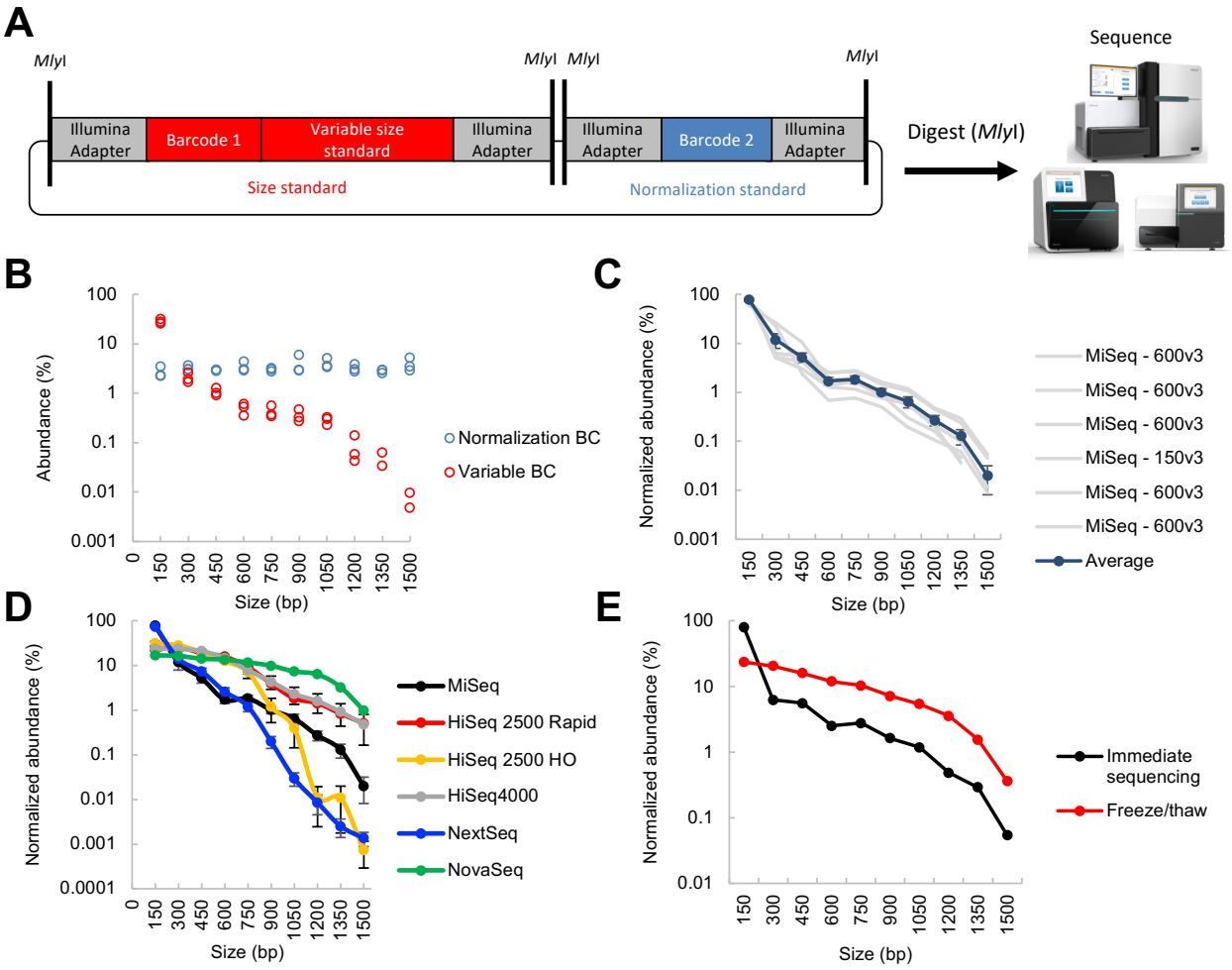

Figure 3. Illumina size standards allow measurement of sequencer-specific size biases. A) Design of REcount-based Illumina size standard constructs. Each standard construct contains a normalization barcode, as well as a barcode associated with a variable size standard that can be liberated by *Mly*I digestion and directly sequenced. B) Raw abundance data for all 30 size standards and normalization barcodes from a MiSeq run. C) Run-to-run variability of multiple MiSeq runs (n = 6 flow cells). D) Size bias profiles of the MiSeq (n=6 flow cells), HiSeq 2500 Rapid (n=1 flow cell, 2 lanes), HiSeq 2500 High Output (HO, n=2 flow cells, 10 lanes), HiSeq 4000 (n=3 flow cells, 6 lanes), NextSeq (n=4 flow cells), and NovaSeq (n=1 flow cell, 1 lane) sequencers. E) Size bias profiles of the same library either clustered on the MiSeq immediately after denaturation, or clustered after freezing and thawing the denatured library. Error bars are +/- s.e.m.

128 previous anecdotal observations (Figure 3D). However, the magnitude of this effect and the
129 shapes of the size bias curves differ substantially between the MiSeq, HiSeq 2500, HiSeq 4000,
130 NextSeq, and NovaSeq (Figure 3D). Differences were also seen between the HiSeq 2500 in
131 Rapid Run (onboard clustering) and High Output (cBot clustering) modes (Figure 3D). In
132 addition, we observed an effect of molecule length on sequencing quality score, with a general
133 trend towards longer molecules having lower quality scores (Supplemental Figure S5). The
134 magnitude of the effect of molecule length on sequence quality varied among the different
135 instruments.

136 The denaturation process can also affect the size bias observed on Illumina instruments.
137 Denatured libraries are sometimes saved for re-sequencing in the case of a run failure (although
138 Illumina's best practices recommend preparing freshly denatured libraries). To test whether
139 freshly denatured libraries perform differently from frozen denatured libraries, we sequenced a
140 freshly denatured library on a MiSeq, and the same denatured library one day later, after a
141 freeze-thaw cycle, on a second MiSeq. The freeze-thaw cycle had a substantial effect on the
142 size bias profile of this library; in particular, there was a dramatic reduction in the fraction of 150
143 bp molecules observed, resulting in a corresponding upward shift of the curve (Figure 3E). It is
144 likely that this shift reflects differential re-annealing of 150 bp fragments (which are in molar
145 excess due to the presence of the large number of similarly sized normalization barcodes), or
146 other small library molecules in the sequencing pool. This observation suggests that some of
147 the difference in clustering size bias observed between the different platforms may be due to
148 differences in denaturation conditions, the amount of time between loading the library and
149 clustering, and whether the clustering process takes place in a chilled compartment (such as on
150 the MiSeq) or not (such as the HiSeq2500 and NextSeq). Consistent with this idea, the variation
151 between HiSeq2500 and HiSeq 4000 flow cells is much larger than the variation between the
152 lanes on the same flow cell (Supplemental Figure S5, Supplemental Figure S6).
153 It is also likely that a portion of the variability between flow cells is due to differences in
154 the size distributions of the libraries being sequenced together with the synthetic size standards,
155 as competition for clustering will occur between all molecules in the sequencing lane. We
156 observed a shift in the curve corresponding to a decreased representation of the larger size
157 standards when they were sequenced together with a library containing a significant amount of
158 material that was smaller than 300 bp on the HiSeq 4000 (Supplemental Figure 6). Although the
159 size standards were sequenced together with different libraries across the different instruments,
160 this context-dependent clustering is not sufficient to explain the large differences we see
161 between different instruments. For example, libraries with similar average sizes and distributions
162 yielded dramatically different measurements of size bias on the NextSeq versus the HiSeq 4000
163 (Supplemental Figure 6).
164 Surprisingly, we also detected an instance of construct-specific size bias, specifically on
165 the HiSeq 2500 platform in Rapid Run mode (Supplemental Figure S5). In contrast to the
166 MiSeq, HiSeq 2500 High Output, HiSeq4000, NextSeq, and NovaSeq where no systematic
167 construct-specific biases were observed, the size bias curves for the 16S, GAPDH, and alpha-
168 Tubulin constructs separated as size increased, with 16S showing much less of a drop-off with
169 increased molecule size. One possible explanation for this difference is that the 16S rRNA gene

170 has substantial secondary structure (Woese et al. 1980), which may serve to shorten the
171 effective length of the molecule during the clustering process. This phenomenon may be due to
172 differences in the clustering process or temperature on this platform, which may be less
173 effective at dissociating the secondary structure of the 16S rRNA gene
174 (https://support.illumina.com/bulletins/2016/10/considerations-when-migrating-nonillumina-
175 libraries-between-sequencing-platforms.html). The HiSeq and MiSeq also have different
176 recommended NaOH concentrations for denaturing libraries. It is possible that long molecules,
177 particularly those with highly stable secondary structure, are incompletely denatured under the
178 HiSeq denaturing conditions.
179
180 **Discussion**
181 In summary, we describe REcount, a novel strategy for obtaining highly accurate and
182 precise PCR-free NGS-based measurements of engineered genetic constructs. Similar
183 constructs could be incorporated into shRNA, CRISPR, and transposon libraries to improve
184 quantification of these molecules in pooled genetic screens. Currently, such measurements are
185 prone to bias introduced by PCR, as we observed for both the BC and V4 amplicons (Figure 1,
186 Supplemental Figure S4). Such sequence-specific amplification biases are often mitigated by
187 including input controls, which are thought to accurately model amplification biases. However,
188 amplification biases can be impacted by template concentration and by the context of the other
189 molecules in the amplification reaction (Gohl et al. 2016), and can limit the sensitivity of these
190 assays by compressing the dynamic range (Supplemental Figure S4). One challenge of
191 employing PCR-free quantification barcodes in these contexts is the large amount of genomic
192 DNA relative to the PCR-free barcode construct. However, we have shown that transposon
193 pools can be quantified from isolated *E. coli* genomic DNA using this approach (data not
194 shown). We further demonstrated that multiplexing of REcount measurements is possible using
195 orthogonal restriction enzymes (Figure 2).
196 We used REcount to measure size bias on several different Illumina sequencers. We
197 found that size bias can vary between runs and instruments and that the denaturation procedure
198 can affect the size bias (Figure 3). Due to the competitive clustering of molecules of different
199 sizes, it is likely that a portion of the variability between runs and lanes is due to differences in
200 the size distributions of the libraries being sequenced together with the synthetic size standards.
201 Thus, the shape of the size bias curve is likely sensitive to both the size distribution of the
202 libraries being sequenced along with the size standards, as well as the proportion of the lane
203 devoted to the size standards.

While such size biases would not be expected to affect randomly sheared libraries, these results indicate that care should be taken when interpreting quantitative measurements or comparing data across different platforms. This is particularly true in cases where library size distributions are non-random such as in several chromatin profiling methods (e.g. ATAC-Seq (Buenrostro et al. 2013), FAIRE-Seq/MAINE-Seq (Ponts et al. 2010)), approaches that use restriction digestion to fragment DNA (e.g. RAD-Seq (Andrews et al. 2016)), amplicons that vary in length (e.g. fungal ITS sequencing (Taylor et al. 2016)), or techniques such as TAIL-Seq (Chang et al. 2014) that explicitly seek to measure molecule length. Constructs such as those described here could be routinely spiked into Illumina sequencing runs to monitor size bias, similar to the use of PhiX to report on sequencing error rates and other base-calling metrics.

214 **Methods**

215 **Synthesis and cloning of REcount plasmids**

216 Even and staggered pool plasmids

217 The plasmids comprising the even and staggered pools were designed to include a portion of
218 the 16S rRNA gene from one of twenty different bacterial species, modeled on the Human
219 Microbiome Project mock microbial communities (HM-276D and HM-277D, (2012a, 2012b), with
220 a 3 bp "TCT" sequence tag added at an analogous position in each construct. These constructs
221 also contained an I-SceI site, allowing for linearization of the plasmids, and a REcount
222 construct, consisting of a unique 20 bp DNA barcode, flanked by Illumina adapters and *Mly*I
223 restriction sites, spaced in a manner to precisely liberate the Illumina adapter-containing
224 barcode construct (Supplemental_File_1). These constructs were synthesized as DNA tiles by
225 SGI-DNA and assembled into full-length constructs using the BioXP 3200 (SGI-DNA). The
226 assembled DNA fragments were A-tailed using the A-tailing module from NEB, cloned into
227 pCR2.1 using a TOPO TA cloning kit (Thermo Scientific), and transformed into OneShot TOP10
228 chemically competent *E. coli* (Thermo Scientific). Multiple colonies were selected, DNA was
229 isolated using a Qiagen MiniPrep kit, and sequence-verified clones were identified by Sanger
230 sequencing with the following primers:
231 M13F: GTAAAACGACGGCCAG
232 M13R: CAGGAAACAGCTATGAC
233 The twenty sequence-verified plasmids were quantified using a Quant-iT PicoGreen dsDNA
234 assay (Thermo Fisher Scientific), normalized to 50 ng/µl, and pooled at equal volume to create
235 the original even pool. The re-pooled even pool and staggered pool were made by adjusting the
236 volume pooled based on the initial PCR-free sequencing data of the original even pool.
237

238 Orthogonal enzyme multiplexing plasmids

239 Four synthetic gene fragments were synthesized (IDT) in the pIDTSmart-Amp plasmid
240 backbone, consisting of an Illumina adapter containing construct with internal *Pac*I and *Pme*I
241 site, and flanked by a pair of either *Mly*I, *Bsm*I, *Bts*$^{\alpha}$I, or *Bsr*DI sites. The full constructs were
242 also flanked by a pair of *Sbf*I sites (Supplemental File 2). In order to make a collection of
243 barcode-containing constructs, the plasmid templates were amplified using the following
244 template-specific primers, and a Golden Gate cloning reaction was used to re-generate the
245 circular plasmid:
246 UMGC_350_MlyI_barcode_p5:
247 NNNNGGTCTCTACTTATCCWWNNNWWNNNAGATCGGAAGAGCGTCGTGTAG

248 UMGC_350_MlyI_barcode_p7:
249 NNNNGGTCTCTAAGTGCAANNNWWNNNWWAGATCGGAAGAGCACACGTCTGAA
250 UMGC_350_BsmI_barcode_p5:
251 NNNNGGTCTCTGGTTATCCNNSSNNSSNNAGATCGGAAGAGCGTCGTGTAG
252 UMGC_350_BsmI_barcode_p7:
253 NNNNGGTCTCTCCAAGCAANNSSNNSSNNAGATCGGAAGAGCACACGTCTGAA
254 UMGC_350_BtsI_barcode_p5:
255 NNNNGGTCTCTGAACATCCNNNWWNNNWWAGATCGGAAGAGCGTCGTGTAG
256 UMGC_350_BtsI_barcode_p7:
257 NNNNGGTCTCTGTTCGCAANNNWWNNNWWAGATCGGAAGAGCACACGTCTGAA
258 UMGC_350_BsrDI_barcode_p5:
259 NNNNGGTCTCTATGAATCCNNSSNNSSNNAGATCGGAAGAGCGTCGTGTAG
260 UMGC_350_BsrDI_barcode_p7:
261 NNNNGGTCTCTTCATGCAANNSSNNSSNNAGATCGGAAGAGCACACGTCTGAA
262 Briefly, PCR reactions were set up using the following recipe: 1 µl plasmid DNA (20 ng/µl), 2.5
263 µl primer 1 (10 µM), 2.5 µl primer 2 (10 µM), 19 µl water, and 25 µl 2x Q5 master mix (NEB).
264 PCR amplification was carried out using the following cycling conditions: 98°C for 30 seconds,
265 followed by 30 cycles of 98°C for 20 seconds, 60°C for 15 seconds, 72°C for 1.5 minutes,
266 followed by 72°C for 5 minutes. Golden Gate reactions(Engler et al. 2008, 2009) were set up
267 using the following recipe:
268 1 µl Barcoding PCR product from above, 2 µl NEB Cutsmart buffer, 2 µl 10mM ATP (NEB), 12.5
269 µl nuclease-free water, 0.5 µl *Bsa*I-HF, 1 µl T4 DNA ligase (NEB 400,000 U/ml), 1 µl *Pac*I.
270 Golden Gate reactions were cycled with the following conditions: 10 cycles of 37°C for 5
271 minutes, 21°C for 5 minutes, then 1 cycle 37°C for 10 minutes, then 1 cycle 80°C for 20
272 minutes. Golden Gate reactions were transformed into chemically competent *E. coli* 5alpha cells
273 (NEB). Colonies were picked and DNA was isolated using a Qiagen Mini-Prep kit. Uniquely
274 barcoded constructs were identified by Sanger sequencing with the following primers:
275 UMGC_350-pIDT-Smart-For: CTGAGGCTCGTCCTGAATGATA
276 UMGC_350-pIDT-Smart-Rev: ACCGATCATACGTATAATGCCGTAA
277 The twelve sequence-verified plasmids were quantified using a Quant-iT PicoGreen dsDNA
278 assay (Thermo Fisher Scientific), normalized to 50 ng/µl, and pooled at equal volume to create
279 the orthogonal enzyme multiplexing test pool. Subsequent NGS analysis indicated that some of
280 these clones were mixed isolates, as other barcodes that had not been detected by Sanger

281 sequencing were present in the NGS data sets. Analysis is based on the Sanger-verified
282 barcodes only.
283
284 Illumina size standard plasmids
285 Illumina size standards were designed using three different template molecules as backbones
286 for the variable length fragment; the 16S rRNA gene (16S) from *E. coli*, the *alpha-Tubulin84B*
287 gene (Tubulin) from *D. melanogaster*, and the *Glyceraldehyde-3-phosphate dehydrogenase 1*
288 (GAPDH) gene from *D. melanogaster* (Supplemental Figure S5). Any naturally occurring *Mly*I
289 sites in these fragments were modified to remove this restriction site. The variable length size
290 standards represent nested fragments of these three genes with breakpoints chosen to
291 generate specific molecule lengths, with GC contents between 40-60% (Figure 3, Supplemental
292 Figure S5). In order to minimize repetitive sequences, different adapters were used for the
293 normalization and variable size standards (Nextera and TruSeq, respectively), and the
294 normalization and size standards were synthesized in opposite orientations in the construct.
295 Both the Illumina adapter flanked variable and normalization barcode constructs were flanked
296 by *Mly*I restriction sites. The Illumina size standard constructs were synthesized by GenScript in
297 the pUC57 cloning vector (Supplemental File 3). Approximately 4 µg of each lyophilized plasmid
298 was resuspended in 40 µl of EB (Qiagen). Plasmids were quantified using a Quant-iT
299 PicoGreen dsDNA assay (Thermo Fisher Scientific) and normalized to 10 nM to account for the
300 variable sizes of the plasmids, then pooled at an equimolar ratio.
301
302 **qPCR validation of ddPCR assays**
303 A set of primers allowing amplification between the construct-specific barcode and the Illumina
304 flow cell adapter, either in the forward orientation (assay 1, where the construct-specific primer
305 was paired with the p7 primer) or reverse orientation (assay 2, where the construct-specific
306 primer was paired with the p5 primer) were designed and synthesized (Integrated DNA
307 Technologies, Supplemental File 4). In order to validate these assays, we performed qPCR
308 amplification of each individual plasmid, the even plasmid pool, and a negative control (water)
309 with each of the 40 primer sets, as well as a p5/p7 positive control (which is expected to amplify
310 all constructs). PCR reactions were set up as follows: 3 µl template DNA (0.05 ng/µl), 1.06 µl
311 nuclease-free water, 0.6 µl 10x Qiagen PCR buffer, 0.24 µl $MgCl_2$ (25 mM), 0.3 µl DMSO, 0.048
312 µl dNTPs (25 mM), 0.12 µl ROX (25 µM), 0.003 µl SYBR (1000x), 0.03 µl Qiagen Taq (5 U/µl),
313 0.3 µl primer 1 (10 µM), and 0.3 µl primer 2 (10 µM). Reactions were amplified on an ABI 7900
314 with the following cycling conditions: 95ºC for 5 minutes, followed by 35 cycles of: 94ºC for 30

315 seconds, 55ºC for 30 seconds, and 72ºC for 30 seconds, followed by incubation at 72ºC for 1
316 minute. For each primer set, Ct values were normalized to the mean Ct for that primer set
317 across all plasmids and plotted as a heatmap (Supplemental Figure S5).
318
319 **ddPCR**
320 The re-pooled even plasmid mix was quantified using a Quant-iT PicoGreen dsDNA assay
321 (Thermo Fisher Scientific), diluted to 1 ng/µl, and further diluted 1:10,000 to bring the pool to the
322 correct concentration for digital quantification. The following ddPCR reactions were prepared: 5
323 µl template DNA, 0.44 µl primer 1 (10 µM), 0.44 µl primer 2 (10 µM), 5.12 µl water, and 11 µl
324 EvaGreen reaction mix (BIO-RAD). In addition, 2 µl of *I-SceI* was added to the ddPCR master
325 mix to linearize the plasmid DNA templates, resulting in between 0.02 and 0.075 µl of *I-SceI* per
326 reaction. Emulsion droplets were generated using a QX200 Droplet Generator (BIO-RAD)
327 following the manufacturer's instructions, transferred to a 96-well PCR plate, and cycled using
328 the following conditions: 95ºC for 10 minutes, followed by 40 cycles of: 95ºC for 30 seconds and
329 55ºC for 1 minute, followed by a final extension step of 72ºC for 5 minutes, and a 12ºC hold.
330 Droplets were counted using a QX200 Droplet Reader (BIO-RAD). The re-pooled even plasmid
331 mix was run in triplicate for both the forward and reverse assays. Single replicates of both the
332 original even pool and the staggered pool were run for both assays. For the staggered pool, the
333 extent of dilution of the 1 ng/µl plasmid pool was varied such that the template abundance of the
334 plasmid targeted by the primer set was expected to be at the correct concentration for digital
335 quantification. Data was analyzed using QuantaSoft Analysis Pro software (BIO-RAD).
336 Replicate measurements were averaged (when available) for both ddPCR assays in order to
337 arrive at a measurement of average ddPCR counts for each construct. Data from the assay was
338 not included in cases where there was not clear separation between positive and negative
339 droplets.
340
341 **Sequencing library preparation**
342 <u>Even and staggered pool REcount measurements</u>
343 The following *Mly*I digests were set up for PCR-free quantification: 200-500 ng even or
344 staggered pool DNA, 2 µl Cutsmart buffer (NEB), 1 µl MlyI (NEB), and volume was adjusted to
345 20 µl with nuclease-free water. Digests were incubated at 37ºC for 1 hour, followed by 20
346 minutes at 65ºC. 30 µl of water was added to each digest (to bring the volume up to 50 µl). 30 µl
347 (0.6x) of AmpureXP beads (Beckman Coulter) were added and after a 5 minute incubation,
348 beads were collected on a magnet and the supernatant was transferred to a new tube

349 (discarded beads). 80 µl (1x) of AmpureXP beads was added, washed 2x for 30 seconds using
350 fresh 80% ethanol, and beads were air dried for 10 minutes, followed by elution in 20 µl of EB
351 (Qiagen). Libraries were quantified using a Quant-iT PicoGreen dsDNA assay (Thermo Fisher
352 Scientific), fragment sizes were assessed using an Agilent Bioanalyzer High Sensitivity assay,
353 and libraries were normalized to 2 nM for sequencing.
354
355 Even and staggered pool PCR-based measurements
356 Barcode construct (BC) library preparation:
357 The following PCR reactions were set up to amplify the BC constructs: 1 µl DNA (1 ng/µl), 5 µl
358 10x Qiagen PCR buffer, 2 µl $MgCl_2$ (25 mM), 2.5 µl DMSO, 0.4 µl dNTPs (25 mM), 0.25 µl
359 Qiagen Taq (5 U/µl), 2.5 µl primer 1 (10 µM), 2.5 µl primer 2 (10 µM), and 33.85 µl nuclease-
360 free water.
361 The following primers were used to amplify the BC constructs:
362 p5: AATGATACGGCGACCACCGA
363 p7: CAAGCAGAAGACGGCATACGA
364 Samples were amplified using the following cycling conditions: 95ºC for 5 minutes, followed by
365 10, 20, 30, or 40 cycles of 94ºC for 30 seconds, 55ºC for 30 seconds, and 72ºC for 30 seconds,
366 followed by incubation at 72ºC for 10 minutes. Libraries were quantified using a Quant-iT
367 PicoGreen dsDNA assay (Thermo Fisher Scientific), fragment sizes were assessed using an
368 Agilent Bioanalyzer High Sensitivity assay, and libraries were normalized to 2 nM for
369 sequencing.
370
371 V4 fragment library preparation:
372 The following PCR reactions were set up in triplicate to amplify the V4 constructs: 2 µl DNA (0.1
373 ng/µl), 0.5 µl primer 1 (10 µM), 0.5 µl primer 2 (10 µM), 2 µl nuclease-free water, and 5 µl 2x Q5
374 master mix. The following primers were used:
375 V4_515F_Nextera:
376 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGCCAGCMGCCGCGGTAA
377 V4_806R_Nextera:
378 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT
379 Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed
380 by 10, 20, 30, or 40 cycles of 98°C for 20 seconds, 55°C for 15 seconds, 72°C for 1 minute,
381 followed by 72°C for 5 minutes.

382 After initial amplification, PCR reactions were diluted 1:60 in nuclease-free water, and used as
383 templates in the following indexing reactions: 3 µl PCR 1 (1:60 dilution), 1 µl indexing primer 1
384 (5 µM), 1 µl indexing primer 2 (5 µM), and 5 µl 2x Q5 master mix. The following indexing primers
385 were used (X indicates the positions of the 8 bp indices):
386 Forward indexing primer:
387 AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC
388 Reverse indexing primer:
389 CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG
390 Reactions were amplified using the following cycling conditions: 98°C for 30 seconds, followed
391 by 10 cycles of 98°C for 20 seconds, 55°C for 15 seconds, 72°C for 1 minute, followed by 72°C
392 for 5 minutes. The full indexing PCR reactions were then purified and normalized using a
393 SequalPrep normalization plate (Thermo Fisher Scientific), followed by elution in 20 µl of elution
394 buffer. An even volume of the normalized libraries was pooled and concentrated using 1x
395 AmpureXP beads (Beckman Coulter). Pooled libraries were quantified using a Qubit dsDNA
396 broad-range assay (Thermo Fisher Scientific), fragment sizes were assessed using an Agilent
397 Bioanalyzer High Sensitivity assay, and libraries were normalized to 2 nM for sequencing.
398
399 Orthogonal enzyme multiplexing tests
400 The twelve-plasmid orthogonal enzyme pool was cut with one of 5 different enzymes (in
401 separate reactions) using the following recipe and enzyme-specific incubation conditions: 20 µl
402 DNA (1 µg), 4 µl NEB buffer (CutSmart or NEB 2.1, depending on enzyme), 2 µl Enzyme (either
403 *Mly*I [37°C for 1 hour, followed by 65°C for 20 minutes], *Bsm*I [65°C for 1 hour, followed by 80°C
404 for 20 minutes], *Bts*$^{\alpha}$I [55°C for 1 hour], or *Bsr*DI [65°C for 1 hour, followed by 80°C for 20
405 minutes], or *Sbf*I [37°C for 1 hour, followed by 80°C for 20 minutes]), and 14 µl water. 10 µl
406 (0.5x) of AmpureXP beads (Beckman Coulter) was added to 20 µl of digested DNA and after a
407 five minute incubation, the beads were collected on a magnet and the supernatant was
408 transferred to new tube (discarded beads). 20 µl of AmpureXP beads was added, and the
409 beads were washed 2x for 30 seconds using fresh 80% ethanol, then air dried for 10 minutes,
410 before eluting in 20 µl of EB (Qiagen). Libraries were quantified using a Quant-iT PicoGreen
411 dsDNA assay (Thermo Fisher Scientific), fragment sizes were assessed using an Agilent
412 Bioanalyzer High Sensitivity assay, and libraries were normalized to 2 nM for sequencing.
413
414 Illumina size standards

415 The following digest of the Illumina size standard pool was set up: 175 µl DNA (10 nM), 20 µl
416 CutSmart buffer (NEB), 5 µl *Mly*I (NEB). The reaction was incubated at 37°C for 1 hour,
417 followed by 65°C for 20 minutes. The library was quantified using a Quant-iT PicoGreen dsDNA
418 assay (Thermo Fisher Scientific), fragment sizes were assessed using an Agilent Bioanalyzer
419 High Sensitivity assay, and libraries were normalized to 2 nM for sequencing.
420
421 **Sequencing**
422 DNA libraries were denatured with NaOH and prepared for sequencing according to the
423 protocols described in the Illumina MiSeq, NextSeq, HiSeq 2500, HiSeq 4000, and NovaSeq
424 Denature and Dilute Libraries Guides. Libraries were generally sequenced along with other
425 samples in a fraction of a sequencing lane.
426
427 **Data analysis**
428 REcount data was analyzed using custom R and Python scripts and BioPython (Cock et al.
429 2009). The first 20 bp of the sequencing reads was mapped against a barcode reference file
430 (Supplemental Files 5-8), with a maximum of 2 mismatches allowed. A generalized script for
431 counting REcount barcodes contained in a reference file is available on Github
432 (https://github.com/darylgohl/REcount). Analysis of the V4 amplicon data was performed using
433 the reference-based mapping pipeline described here: https://bitbucket.org/jgarbe/gopher-
434 pipelines/wiki/metagenomics-pipeline.rst, using the reference file in Supplemental File 9 to build
435 the bowtie2 index (Langmead and Salzberg 2012). For the analysis of quality scores
436 (Supplemental Figure S5), the data for all runs on a given platform was concatenated into a
437 single fastq file, the split into individual fastq files for each individual construct, based on the 20
438 bp sequence barcodes in each construct. Next, the reads were trimmed to 50 bp using cutadapt
439 (Martin 2011), so that all constructs and sequencing runs could be compared in a standardized
440 manner. Mean quality scores were calculated for each construct that was represented by at
441 least 100 reads in the data set.
442
443 **Data access**
444 Sequencing data files are available through the NCBI Sequence Read Archive (BioProject:
445 PRJNA431017).
446
447

Acknowledgements

We thank the staff of the University of Minnesota Genomics Center for advice, technical support, and data generation, and the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley for data generation. We also thank Alessandro Magli, Daniel Schmidt, Igor Libourel, Steve Bowden, Benjamin Auch, John Garbe, and Nagendra Palani for helpful discussions. This work was supported by a grant from the University of Minnesota-Mayo Translational Product Development Fund to D.M.G. and K.B.B. (National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR000114). The Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley was supported by an NIH S10 OD018174 Instrumentation Grant.

Author contributions

D.M.G. and K.B.B. conceived and designed the experiments, analyzed data, and wrote the manuscript. D.M.G., A.B., D.J., S.A., B.B., and S.M. conducted the experiments.

Disclosure declaration

The REcount PCR-free quantification barcode technology described here is included in US patent application numbers 62/332,879, 62/630,463, and PCT/US17/31271. D.M.G. is a shareholder and CSO of CoreBiome, Inc. K.B.B. is a shareholder and COO of CoreBiome, Inc.

References

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **17**: 81–92.

Bhang HC, Ruddy DA, Krishnamurthy Radhakrishna V, Caushi JX, Zhao R, Hims MM, Singh AP, Kao I, Rakiec D, Shaw P, et al. 2015. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* **21**: 440–448.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

Chang H, Lim J, Ha M, Kim VN. 2014. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Mol Cell* **53**: 1044–52.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–3.

Engler C, Gruetzner R, Kandzia R, Marillonnet S. 2009. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes ed. J. Peccoud. *PLoS One* **4**: e5553.

Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. ed. H.A. El-Shemy. *PLoS One* **3**: e3647.

Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, et al. 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **26**: 317–325.

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, et al. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* **34**: 942–949.

Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, et al. 2011. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. *Anal Chem* **83**: 8604–8610.

Illumina. 2014. *Nextera® Library Validation and Cluster Density Optimization*.

Kebschull JM, Garcia da Silva P, Reid AP, Peikon ID, Albeanu DF, Zador AM. 2016. High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron* **91**: 975–987.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**: 9530–5.

Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2011. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**: 72.

Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32**: 267–273.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. 2016. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science (80-)* **353**: aaf7907.

Peikon ID, Kebschull JM, Vagin V V, Ravens DI, Brouzes E, Corrêa IR, Bressan D, Zador A. 2017. Using high-throughput barcode sequencing to efficiently map connectomes. *bioRxiv* 099093.

Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, Hardiman G, Lonardi S, Le Roch KG. 2010. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res* **20**: 228–38.

Rodriguez-Barrueco R, Marshall N, Silva JM. 2013. Pooled shRNA screenings: experimental approach. *Methods Mol Biol* **980**: 353–70.

Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, et al. 2014. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (80-)* **343**: 84–87.

Shalem O, Sanjana NE, Zhang F. 2015. High-throughput functional genomics using CRISPR–Cas9. *Nat Rev Genet* **16**: 299–311.

Sims D, Mendes-Pereira AM, Frankum J, Burgess D, Cerone M-A, Lombardelli C, Mitsopoulos C, Hakas J, Murugaesu N, Isacke CM, et al. 2011. High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol* **12**: R104.

Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C. 2009. Quantitative phenotyping via deep barcode sequencing. *Genome Res* **19**: 1836–42.

Strezoska Ž, Licon A, Haimes J, Spayd KJ, Patel KM, Sullivan K, Jastrzebski K, Simpson KJ, Leake D, van Brabant Smith A, et al. 2012. Optimized PCR conditions and increased shRNA fold representation improve reproducibility of pooled shRNA screens. ed. K.T. Jeang. *PLoS One* **7**: e42341.

Taylor DL, Walters WA, Lennon NJ, Bochicchio J, Krohn A, Caporaso JG, Pennanen T. 2016. Accurate Estimation of Fungal Diversity and Abundance through Improved Lineage-Specific Primers Optimized for Illumina Amplicon Sequencing. *Appl Environ Microbiol* **82**: 7217–7226.

van Opijnen T, Camilli A. 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* **11**: 435–442.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science (80-)* **343**: 80–84.

Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius R, Gutell R, Hogan JJ, et al. 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* **8**: 2275–2294.

2012a. A framework for human microbiome research. *Nature* **486**: 215–21.

2012b. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* **7**: e39315.