

1 Reliable Multiplex Sequencing with Rare Index

2 Mis-Assignment on DNB-Based NGS Platform

3 Qiaoling Li^{1,2,✉}, Xia Zhao^{1,2,✉}, Wenwei Zhang^{1,2,✉}, Lin Wang³, Jingjing Wang^{1,2}, Dongyang

4 Xu^{1,2}, Zhiying Mei⁴, Qiang Liu⁵, Shiyi Du⁴, Zhanqing Li^{1,2}, Xinming Liang⁴, Xiaman Wang⁵,

5 Hanmin Wei⁴, Pengjuan Liu^{1,2}, Jing Zou⁴, Hanjie Shen^{1,2}, Ao Chen^{1,2}, Snezana Drmanac^{1,3}, Jia

6 Sophie Liu³, Li Li^{1,2}, Hui Jiang⁴, Yongwei Zhang^{1,3}, Jian Wang^{1,6}, Huanming Yang^{1,6}, Xun Xu^{1,2},

7 Radoje Drmanac^{1,2,3,4,*}, Yuan Jiang^{3,*}

8

9 ¹ BGI-Shenzhen, Shenzhen 518083, China.

10 ² China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China.

11 ³ Advanced Genomics Technology Lab, Complete Genomics Inc., 2904 Orchard Pkwy, San Jose,
12 California 95134, USA.

13 ⁴ MGI, BGI-Shenzhen, Shenzhen 518083, China.

14 ⁵ BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China.

15 ⁶ James D. Watson Institute of Genome Sciences, Hangzhou 310058, China.

16 [✉]: contributed equally to this work.

17 ^{*}: corresponding authors. yjiang@completegenomics.com and

18 rdrmanac@completegenomics.com

19

20

21 **Abstract**

22 **Background**

23 Massively-parallel-sequencing, coupled with sample multiplexing, has made
24 genetic tests broadly affordable. However, intractable index mis-assignments
25 (commonly exceeds 1%) were repeatedly reported on some widely used sequencing
26 platforms.

27 **Results**

28 Here, we investigated this quality issue on BGI sequencers using three library
29 preparation methods: whole genome sequencing (WGS) with PCR, PCR-free WGS,
30 and two-step targeted PCR. BGI's sequencers utilize a unique DNB technology which
31 uses rolling circle replication for DNA-nanoball preparation; this linear amplification is
32 PCR free and can avoid error accumulation. We demonstrated that single index mis-
33 assignment from free indexed oligos occurs at a rate of one in 36 million reads,
34 suggesting virtually no index hopping during DNB creation and arraying. Furthermore,
35 the DNB-based NGS libraries have achieved an unprecedentedly low sample-to-sample
36 mis-assignment rate of 0.0001% to 0.0004% under recommended procedures.

37 **Conclusions**

38 Single indexing with DNB technology provides a simple but effective method for
39 sensitive genetic assays with large sample numbers.

40

41

42 **Background**

43 NGS technology, with its remarkable throughput and rapidly reduced sequencing
44 cost in the current “Big Data” era, is advancing into clinical practice faster than expected
45 by Moore’s Law. Updated sequencers, such as Illumina’s HiSeq and NovaSeq and
46 BGI’s BGISEQ and MGISEQ, are capable of producing hundreds of gigabases to a few
47 terabases of sequencing data in a single run. Different sequencing platforms share a
48 basic NGS workflow, which includes sample/library preparation (nucleic acid isolation,
49 end repair, size selection, adapter addition, and optional PCR amplification),
50 sequencing (quality control of the library, DNA cluster/array generation, and instrument
51 operation), and data analysis (quality control, data pipeline analysis, and data
52 interpretation)[1, 2]. One of the most common strategies for maximizing efficiency is the
53 multiplexing of samples; a unique index is appended to each sample, and multiple
54 samples are pooled together for sequencing in the same run. After sequencing the
55 library pool including the indexes, each read would then be reassigned to its
56 corresponding sample according to the unique index sequence. This sample
57 multiplexing occurs during library preparation, and indexes can be embedded in DNA
58 constructs in two distinct ways—through ligation using indexed adapters or through
59 PCR amplification using indexed primers.

60 However, researchers must be very careful when analyzing de-multiplexed data
61 because index mis-assignment from multiplexing affects data quality and may lead to
62 false conclusions. Index switching can be introduced during many stages of the library
63 preparation and sequencing and post-sequencing processes, including oligo
64 manufacture error or contamination, reagent contamination during experimental

65 handling, template switching during PCR amplification (recombinant PCR), sequencing
66 artifacts or errors, and bioinformatic errors. For example, Illumina's platforms, especially
67 the ones using the new Illumina clustering chemistry, ExAmp, were reported by different
68 labs to have a total contamination rate of 1% to 7% using dual-indexed adapters[3-5].
69 Although the results would be unaffected or only minimally affected for users who follow
70 the best practices suggested from Illumina's white paper, sequencing to detect low-
71 frequency alleles such as in liquid biopsy or tumor exome sequencing[6], or single cell
72 sequencing[4] could be seriously impacted with single or regular combinatorial dual
73 indexing[3, 5].

74 Here, we demonstrate that using the PCR-free DNA array preparation and
75 sequencing technology of DNB nanoarrays with optimized library preparation protocols
76 and index quality filters, BGI sequencers even with single indexing are practically free
77 from index switching. We observed nearly zero index hopping from free indexes and an
78 individual sample-to-sample leakage rate in each sequencing lane less than 0.0004%.
79 The total index contamination rate was also orders of magnitude lower than the reported
80 index hopping rate on Illumina's sequencers.

81

82 **Results**

83 **High indexing fidelity expected for DNA nanoball technology**

84 BGISEQ platforms load DNBs onto patterned arrays and utilize combinatorial
85 Probe Anchor Synthesis (cPAS) for sequencing[7]. The unique DNB technology

86 employs Phi29 polymerase, which has strong strand displacement activity, and the
87 rolling circle replication (RCR) process to enable linear amplification; each amplification
88 cycle remains independent by using the original circular (single-stranded circle)
89 template (**Fig. 1a**). Therefore, even if errors such as index hopping from incorrectly
90 indexed oligos occur, the false copies will not accumulate. Correct sequences would
91 always be replicated in later DNA copies to ensure the highest amplification fidelity.
92 Thus, we hypothesize that the index hopping should be efficiently prevented on BGI
93 sequencers. To test this hypothesis, we first analyzed two important controls.

94 **Index mis-assignment in controls**

95 The standard WGS library construction method for BGISEQ-500 includes the
96 following major steps: 1) DNA fragmentation, 2) end repair and A-tailing, 3) indexed
97 adapter ligation, 4) PCR amplification, 5) single-stranded circle (ssCir) formation, and 6)
98 DNB preparation (**Fig. 2a**). We introduce unique single indexes into every sample
99 during adapter ligation. Each sample is handled separately until samples are pooled,
100 which is known as multiplexing.

101 To determine whether BGISEQ-500 sequencing accuracy is affected by index
102 hopping, as occurs with Illumina's sequencers [3, 4, 8-11], we examined the rate of
103 index mis-assignment in BGISEQ-500 runs. We ligated eight unique single indexes to
104 eight gene regions, respectively (indexes 1-8) (Supplementary **Table 1**) or to eight
105 water controls lacking DNA inputs (indexes 33-40), and we pooled equal volumes of all
106 samples after PCR amplification. For base positional balance on sequencers, balancing
107 WGS library controls with indexes 41-48 were added at an equal molar ratio prior to

Table 1. Observed frequencies of read mis-assignment in controls.

Experiments	Mis-assignment causes	Index #	Total reads mapped to 8 gene regions			Mis-assignment rate per index
			repeat 1	repeat 2	repeat 3	
Experimental groups	N.A.	Barcode 1-8	41686373	44974964	42874988	N.A.
Empty controls	Physical barcode hopping	Barcode 33-40	9	14	6	1 in 36 million reads
Balancing library controls	Total mis-assignments occur after ssCir	Barcode 41-48	612	650	724	1 in 0.5 million reads
All groups	All above	All indexes above	41686994	44975628	42875718	N.A.

Experimental groups, WGS-like libraries prepared separately using indexes 1 to 8; empty controls, indexes 33-40 and reagents used but without sample DNA; balancing library controls, samples prepared and indexed with indexes 41-48 independently and pooled with test samples after ssCir formation; all groups, total reads of all the indexes. Reads were presented after applying a Q30>60% filter.

108 DNB preparation (see Methods). To avoid index mis-assignments from oligo synthesis
 109 contamination, we ordered indexes 1-8 from IDT (U.S.) and indexes 33-48 from
 110 Invitrogen (China) using their regular synthesis services.

111 The results of assessing different index mis-assignments on BGISEQ-500 are
 112 shown in **Table 1**. All reads passing a quality filter (Q30>60%) were de-multiplexed with
 113 perfect matches on the index regions before mapping to the eight gene regions. Indexes
 114 33-40 were used in empty controls lacking sample DNA. The physical index hopping of
 115 the free indexed oligos for all eight indexes occurred at a rate of 2.16E-07 (9 out of
 116 41,686,994), 3.11E-07 (14 out of 44,975,628), and 1.40E-07 (6 out of 42,875,718) in
 117 three repeats (**Table 1**). In other words, the average per-index probability of this type of
 118 index mis-assignment using the DNB platform is 1 in 36 million reads. This number
 119 does not exclude index contamination in the experimental handling of indexed oligos,
 120 confirming no physical index hopping as we hypothesized.

121 In another control group, balancing libraries of indexes 41-48 were pooled with
 122 experimental samples after ssCir formation and prior to the DNB construction process.

123 The average mis-assignment rate from this control group was 1.92E-06 (<0.0002%, 1 in
124 500,000) per index (total reads with indexes 41-48 mapped to genes 1-8 divided by the
125 total reads of all indexes and then divided by 8). When a Q30>80% filter was applied to
126 remove more low-quality indexes, we found one mismatched read per million mapped
127 reads per index (data not shown). These rare index mis-assignments from balancing
128 library controls represent all mis-assignments that occurred after the single-stranded
129 circles formation step, which includes index hopping during DNB creation, sequencing
130 or bioinformatic errors, and other mis-assignments during DNB sequencing.

131 These controls demonstrated that the BGISEQ platform suffers practically no
132 index hopping from excess free indexed oligos and exceptionally low total mis-
133 assignments from the DNB arraying and sequencing processes. In contrast, Costello M.
134 et al. recently reported index hopping rates of 1.31% and 3.20% for i7 and i5 adapters
135 respectively between a human and an *E.coli* library using Illumina's ExAmp chemistry[5].
136 Furthermore, 689,363 reads resulted from uncorrectable double index switching in a
137 total of 842,853,260 mapped reads. Therefore, i7 and i5 were both swapped in the same
138 DNA, causing sample-to-sample mis-assignment at a rate of 0.08%
139 (689,363/842,853,260), or 1 mis-assignment in 1223 reads. The switching mainly
140 originates from index hopping during ExAmp reactions as their empirical data suggested
141 and results in part from oligo synthesis, handling contamination, or index misreading.

142 Higher contamination from balancing library controls (indexes 41-48) compared
143 with empty controls (indexes 33-40) suggests that there are some other mechanisms of
144 mis-assignment in DNB sequencing process independent of the physical hopping of

145 free indexed oligos. We further investigated these mechanisms to optimize our library
146 preparation protocol and minimize sample barcode mis-assignments.

147 **Index mis-assignment rates for “standard PCR-based WGS”-like
148 libraries**

149 To pinpoint an optimal step for sample pooling, we compared the contamination
150 rates of pooling at different processing steps for indexes 1-8 (**Fig. 2a, Fig. 3a**). Each
151 experimental method was repeated in triplicate; therefore, a total of fifteen multiplexed
152 libraries were loaded and sequenced on fifteen lanes of BGISEQ-500.

153 The overall sequencing quality among all libraries was consistently good, and the
154 mean Q30 score is 91.80%. Before mapping, we de-multiplexed the reads based on
155 their individual indexes allowing for a 1-bp mismatch. The splitting rates were quite
156 uniform among the eight indexes if pooling occurred after PCR amplification. An
157 example of the index split rate for PCR-pooled libraries is shown in **Fig. 3b**. We next
158 mapped all reads to the reference genome, and the mapping rates were 99.20% on
159 average. The read numbers of eight gene regions were counted and **Fig. 3c** shows an
160 example of the read counts mapped for each index at each gene region. The total index
161 contamination was calculated by dividing the sum of all hopped reads by the total reads
162 of all the indexes.

163 The total index contamination rates, implying index hopping of the sequencing
164 lane among indexes 1 to 8, were summarized in **Fig. 3a** for each pooling scenario; the
165 number dropped significantly from 2.6792% with one bead purification (Ad-1B group) to
166 0.1365% when an additional step of bead purification (Ad-2B group) was included to

167 further remove excess adapter oligos after adapter ligation (**Fig. 3a, Supplementary**
168 **Table 2**). The effect of template switching on index contamination can be further
169 eliminated by pooling after PCR amplification. Therefore, the rate was reduced by an
170 additional 7-fold, to 0.0183% (PCR group in **Fig. 3a**), if samples were pooled after PCR
171 amplification. Libraries pooled after DNB formation demonstrated a total contamination
172 rate less than 0.015% (DNB group in **Fig. 3a**). However, pooling after ssCir or DNB
173 formation would slightly increase labor and cost. Taking all of the above into
174 consideration, we conclude that pooling after PCR amplification is optimal to achieve
175 low index contamination.

176 **Explaining and reducing the observed index mis-assignment**

177 Index contamination can be introduced through experimental handling, PCR
178 errors, sequencing errors, oligo synthesis errors, or arraying/clustering methods. We
179 therefore investigated some of these potential causes of the index mis-assignment
180 using the triplicate libraries pooled after PCR in **Fig.3a**. First, each mismatch from index
181 1 to index 8 was retraced to the corresponding DNB and analyzed for sequencing
182 quality. These mismatched DNBs exhibited slightly lower quality scores (average
183 Q30=79.24%) at the genomic region compared with those of the DNBs with correctly
184 assigned indexes (average Q30=89.11%). However, the average Q30 of the index
185 region on mismatched DNBs was only 36.66%, which is significantly lower than that of
186 the index region for the correctly matched DNBs (average Q30=91.19%). These
187 analytical results suggested that in these rare cases in which the true index was not
188 detected, a low-quality false index was assigned. We further questioned whether the

189 mis-assignment in this scenario occurred due to signal bleeding from neighboring DNBs
190 to the affected DNBs. We retraced the positions of DNBs on a chip and calculated the
191 percentage of DNBs that shared the same index sequence with at least one of their four
192 surrounding DNBs. On average, 20.21% of correctly assigned DNBs shared the same
193 index sequence with their neighboring DNBs; however, this percentage was 57.04% for
194 mis-assigned DNBs (data not shown). This result suggested that signal bleeding caused
195 barcode mis-assignment in DNBs that had non-detectable true index signals.
196 Nevertheless, most of these mis-assignments can be adequately removed by
197 implementing a Q30 filter; the total contamination rate of indexes 1-8 dropped from
198 0.0188% to 0.0097% and the average sample-to-sample mis-assignment rate dropped
199 to 0.0001% after applying a Q30>60% filter for these PCR-pooled libraries (**Fig. 3c**).

200 Second, we observed in every run that a higher percentage of reads, especially
201 EFEMP2 and LOX, were mistakenly reassigned to index 7 (highlighted in yellow in **Fig.**
202 **3c**). Through thorough investigation, we found that the majority of these EFEMP2/LOX
203 reads mis-assigned to index 7 were perfectly matched and that the quality was high at
204 the index region (average Q30=85.03% and 82.38%, respectively). However, the
205 hamming distance between indexes 2 and 7 is 8, and the hamming distance between
206 indexes 3 and 7 is 9; therefore, the exceptionally highly contaminated EFEMP2/LOX
207 reads even with the Q30>60% filter were less likely to be caused by random sequencing
208 errors. Indexed oligos in this experiment were ordered using IDT's regular oligo
209 synthesis pipeline instead of TruGrade oligo synthesis, which is specifically advertised
210 for NGS. It is highly likely that the index 7 oligo contaminated all other oligos during
211 synthesis or oligo handling. Because reads of index 7 consisted of both correct and

212 false reads that cannot be differentiated, we excluded data from index 7, which reduced
213 the total contamination rate from 0.0183% (PCR group in **Fig. 3a**) to only 0.0124% (**Fig.**
214 **4, Supplementary Table 3**). The rate is further reduced by 275%, to 0.0045%, after
215 applying the Q30>60% filter, whereas the percentage of total reads only dropped by 4%
216 (**Fig. 4, Supplementary Table 3**). This evidence suggested that oligo synthesis
217 contamination was another major cause of index mis-assignment in this experiment.
218 The average individual index contamination rate is approximately 1-2 reads/million after
219 removing low-quality reads and oligo contamination (**Fig. 3c**, data not shown).

220

221 **Contamination rate of PCR-free library construction pipeline**

222 In addition to the aforementioned WGS-like library preparation method, a PCR-
223 free workflow is also commonly used in real-world NGS applications such as PCR-free
224 WGS libraries. Another example is BGI's SeqHPV genotyping assay, which utilizes
225 targeted PCR amplification to first enrich the L1 capsid gene region of human
226 papillomavirus (HPV) and then uses a PCR-free protocol for library preparation (**Fig.**
227 **2b**). To determine whether our rare contamination rate is sustained when the PCR-free
228 library preparation pipeline is used, we evaluated the SeqHPV protocol with six HPV-
229 positive control samples on the BGISEQ-500.

230 The 6 positive samples along with 62 negative samples with YH genome (an
231 Asian male diploid genome) and 4 water controls were individually amplified with unique
232 sample indexes (**Table 2a**). Twelve samples from the same row were pooled together

Table 2. Level of contamination for PCR-free library on BGISEQ-500.

a. Sample arrangement of PCR-free library (HPV).

Template Sample index	YH-1 <i>MGIP-1</i>	HPV11 + YH <i>MGIP-2</i>	YH-1 <i>MGIP-3</i>	YH-1 <i>MGIP-4</i>	YH-1 <i>MGIP-5</i>	YH-1 <i>MGIP-6</i>	H2O-1 <i>MGIP-7</i>	YH-1 <i>MGIP-8</i>	YH-1 <i>MGIP-9</i>	YH-1 <i>MGIP-10</i>	YH-1 <i>MGIP-11</i>	YH-1 <i>MGIP-12</i>	Barcode 1
Template Sample index	YH-2 <i>MGIP-13</i>	YH-2 <i>MGIP-14</i>	H2O-2 <i>MGIP-15</i>	YH-2 <i>MGIP-16</i>	YH-2 <i>MGIP-17</i>	YH-2 <i>MGIP-18</i>	YH-2 <i>MGIP-19</i>	YH-2 <i>MGIP-20</i>	YH-2 <i>MGIP-21</i>	HPV18 + YH <i>MGIP-22</i>	YH-2 <i>MGIP-23</i>	YH-2 <i>MGIP-24</i>	Barcode 2
Template Sample index	YH-3 <i>MGIP-25</i>	YH-3 <i>MGIP-26</i>	YH-3 <i>MGIP-27</i>	YH-3 <i>MGIP-28</i>	YH-3 <i>MGIP-29</i>	YH-3 <i>MGIP-30</i>	YH-3 <i>MGIP-31</i>	YH-3 <i>MGIP-32</i>	YH-3 <i>MGIP-33</i>	YH-3 <i>MGIP-34</i>	YH-3 <i>MGIP-35</i>	YH-3 <i>MGIP-36</i>	Barcode 3
Template Sample index	YH-4 <i>MGIP-37</i>	YH-4 <i>MGIP-38</i>	YH-4 <i>MGIP-39</i>	YH-4 <i>MGIP-40</i>	YH-4 <i>MGIP-41</i>	YH-4 <i>MGIP-42</i>	HPV33 + YH <i>MGIP-43</i>	YH-4 <i>MGIP-44</i>	YH-4 <i>MGIP-45</i>	YH-4 <i>MGIP-46</i>	YH-4 <i>MGIP-47</i>	YH-4 <i>MGIP-48</i>	Barcode 4
Template Sample index	HPV52 + YH <i>MGIP-49</i>	YH-5 <i>MGIP-50</i>	YH-5 <i>MGIP-51</i>	YH-5 <i>MGIP-52</i>	YH-5 <i>MGIP-53</i>	YH-5 <i>MGIP-54</i>	YH-5 <i>MGIP-55</i>	YH-5 <i>MGIP-56</i>	YH-5 <i>MGIP-57</i>	YH-5 <i>MGIP-58</i>	YH-5 <i>MGIP-59</i>	YH-5 <i>MGIP-60</i>	Barcode 5
Template Sample index	YH-6 <i>MGIP-61</i>	YH-6 <i>MGIP-62</i>	YH-6 <i>MGIP-63</i>	YH-6 <i>MGIP-64</i>	YH-6 <i>MGIP-65</i>	YH-6 <i>MGIP-66</i>	H2O-6 <i>MGIP-67</i>	HPV45+11 + YH <i>MGIP-68</i>	YH-6 <i>MGIP-69</i>	YH-6 <i>MGIP-70</i>	YH-6 <i>MGIP-71</i>	YH-6 <i>MGIP-72</i>	Barcode 6

b. Performance of SeqHPV.

Library Index	Sample Index	Total Reads	Mapped Reads	Mapped Rate	Major Types	Information of Major Types	All Information of Types	HBB Score (0-10)	HPV Score (0-10)
1	MGIP002	2470768	1800287	72.90%	HPV11,HBB	HPV11(1348689,14750.9,74.9%);HBB(451597 9833.9,25.1%)	HPV11(1348689,14750.9,74.9%);HBB(451597,9833.9,25.1%)	10	10
2	MGIP022	2653747	2526477	95.20%	HPV18,HBB	HPV18(2309693,8458.3,91.4%);HBB(216783 8458.3,8.6%)	HPV18(2309693,8458.3,91.4%);HBB(216783,8458.3,8.6%)	10	10
3	MGIP029	1793620	1690665	94.30%	HPV31,HBB	HPV31(1566415,81119.5,92.7%);HBB(124250, 5413.0,7.3%)	HPV31(1566415,81119.5,92.7%);HBB(124250,5413.0,7.3%)	10	10
4	MGIP043	1511740	1210189	80.10%	HPV33,HBB	HPV33(940264,3842.6,77.7%);HBB(269904,7 685.1,22.3%)	HPV33(940264,3842.6,77.7%);HBB(269904,7685.1,22.3%)	10	10
5	MGIP049	1641545	1447782	88.20%	HPV52,HBB	HPV52(1236757,7313.3,85.4%);HBB(211023, 7313.3,14.6%)	HPV52(1236757,7313.3,85.4%);HBB(211023,7313.3,14.6%)	10	10
6	MGIP069	2800830	1942883	69.40%	HPV45,HPV11 HBB	HPV45(1497649,6782.4,77.1%);HPV11(25333 7,10173.6,13.0%);HBB(191896,6782.4,9.9%)	HPV45(1497649,6782.4,77.1%);HPV11(25333,10173.6,1 3.0%);HBB(191896,6782.4,9.9%);HPV71(1,6782.4,0.0%)	10	10
8	MGIP002 MGIP029 MGIP049 MGIP069	8 4 17 11	4 75.00% HPV31 HPV52 HPV45,HBB	50.00% HPV11,HBB HPV31(3,0.2,100.0%) HPV52(16,0.2,100.0%) HPV45(5,0.2,71.4%);HBB(2,0.2,28.6%)	HPV11(3,0.2,75.0%);HBB(1,0.2,25.0%) HPV31(3,0.2,100.0%) HPV52(16,0.2,100.0%) HPV45(5,0.2,71.4%);HBB(2,0.2,28.6%)	HPV11(3,0.2,75.0%);HBB(1,0.2,25.0%) HPV31(3,0.2,100.0%) HPV52(16,0.2,100.0%) HPV45(5,0.2,71.4%);HBB(2,0.2,28.6%)	5 0 0 10	10 0 0 10	

c. Index contamination rate of PCR-free libraries.

	Library index	HBB	HPV11	HPV18	HPV31	HPV33	HPV52	HPV45
Read depth	1	2994608	1348826	83	36	14	23	33
	2	2722311	75	2310955	31	17	24	31
	3	1891540	53	65	1566954	10	8	18
	4	2936888	54	90	80	940365	18	25
	5	2289158	61	52	24	14	1237126	22
	6	1747934	253390	53	17	9	18	1497716
	8	27	3	0	3	0	16	5
			14.7309%	0.0009%	0.0004%	0.0002%	0.0003%	0.0004%
Percent of read depth	1		0.0008%	25.2386%	0.0003%	0.0002%	0.0003%	0.0003%
	2			0.0006%	0.0007%	17.1132%	0.0001%	0.0001%
	3						0.0002%	
	4					10.2700%	0.0002%	0.0003%
	5						13.5110%	0.0002%
	6			2.7673%	0.0006%	0.0002%	0.0001%	16.3570%
	8				0.0000%	0.0000%	0.0000%	0.0002%
								0.0001%

a. Positive samples are in ***italic bold***, negative samples with YH genome only are in black font, water controls are ****bolded**** and sample index are in *italic*. Index 7 data was excluded due to its oligo synthesis contamination. **c.** ***Italic bold***, proper combinations; *italic*, improper combinations. The average sample-to-sample mis-assignment rate is 0.0004% without any filtering.

233 after PCR amplification, and then they were ligated with a unique library index (**Table 2a**,
234 **Fig. 2b**). Two empty controls without PCR amplicons were included in the ligation;
235 these were separately tagged by index 7 or 8. The eight libraries were mixed together
236 after ssCir formation and were then subjected to sequencing. After demultiplexing with

237 perfect matches to designed barcodes, BGI's HPV panel precisely detected all six
238 positive samples without any false positive or false negative calls (**Table 2b**). In our
239 assay, we applied quality controls starting from the targeted PCR step, during which
240 four water controls were used to reveal potential sample contamination during PCR
241 amplification. Reads in the water controls were near zero, suggesting no contamination
242 from targeted PCR (**Supplementary Table 4**). When calculating contamination rates for
243 empty controls, we excluded index 7 because of its oligo synthesis contamination as
244 discussed above. Consistent with our previous findings, the empty control, index 8, had
245 only 0.0002% leakage (27 out of 14,582,466) from all of the *HBB* reads (**Table 2c**). This
246 99.9998% precision without any Q30 filter confirms again that the DNB preparation and
247 arraying strategy can minimize index contamination to a great extent. Similar to the
248 WGS library above, the individual sample-to-sample contamination rate was
249 approximately 4 reads/million on average. The total PCR-free library index
250 contamination rate is as low as 0.0118% without any filtering (**Table 2c**).

251 **Contamination rate of two-step PCR library preparation approach**

252 A third popularly used NGS library preparation technique is to embed an index
253 during PCR amplification, as is the case with the BGI lung cancer kit (**Fig. 2c**). The
254 libraries were constructed with index 1 associated with negative control YH DNA, index
255 2 associated with an EGFR L858R mutation at 1%, index 3 associated with a KRAS
256 G12D mutation at 10%, and index 4 associated with an EGFR exon 19 deletion at 50%.
257 NRAS(p.Q61H) is one of the cancer COSMIC sites included in the kit and is used here
258 as a negative control. The mapping rate and capture rate are both greater than 98%,

259 and the uniformity is above 90% (data not shown). We employed unique identifiers
260 (UIDs) to correct and remove PCR and sequencing errors[12, 13]. Before the removal of
261 duplications using UIDs, index contamination existed at ratios from 0.000% to 0.05%
262 (mutant reads divided by the sum of mutant reads and reference reads), but all of these
263 were called “negative” after bioinformatics analysis (**Table 3a**). Moreover, most of the
264 mis-identified reads dropped to 0 after duplication removal, especially for EGFR
265 mutants (**Table 3b**). A 1% sensitivity for mutation detection was demonstrated in this
266 study. Taken together, the BGI lung cancer kit verifies that single indexing on DNB
267 sequencing platforms is not susceptible to read mis-assignment and that it can be used
268 for the precise detection of low-frequency somatic variations such as in cancer.

Table 3. Contamination rate of PCR-introduced adapter library preparation method using MGI lung cancer kit.

a. Contamination rate before removing duplication.

Index	Repeats	EGFR (L858R)			KRAS (G12D)			EGFR (19del)			NRAS (p.Q61H)		
		Reference reads	Mut reads	Mut allele rate	Reference reads	Mut reads	Mut allele rate	Reference reads	Mut reads	Mut allele rate	Reference reads	Mut reads	Mut allele rate
1	Repeat 1	1423408	4	negative	52589	34	negative	31150	0	negative	188086	0	negative
	Repeat 2	1158060	4	negative	54331	33	negative	31047	0	negative	201147	0	negative
2	Repeat 1	1346831	17200	1.2610%	59590	39	negative	40077	0	negative	205321	0	negative
	Repeat 2	1148168	11231	0.9687%	57175	27	negative	36381	0	negative	192472	0	negative
3	Repeat 1	1604176	6	negative	53555	7713	12.5890%	32294	0	negative	199296	2	negative
	Repeat 2	1430975	5	negative	54029	7296	11.8973%	36961	0	negative	200989	4	negative
4	Repeat 1	1321771	3	negative	56766	20	negative	22370	9038	28.7761%	150478	0	negative
	Repeat 2	1275573	7	negative	59610	31	negative	22914	9660	29.6556%	204544	0	negative

b. Contamination rate after removing duplication.

Index	Repeats	EGFR (L858R)			KRAS (G12D)			EGFR (19del)			NRAS (p.Q61H)		
		Reference templates	Mut templates	Mut allele rate	Reference templates	Mut templates	Mut allele rate	Reference templates	Mut templates	Mut allele rate	Reference templates	Mut templates	Mut allele rate
1	Repeat 1	26824	0	negative	6889	2	negative	5295	0	negative	10798	0	negative
	Repeat 2	21904	0	negative	6209	1	negative	5088	0	negative	9617	0	negative
2	Repeat 1	24550	324	1.3026%	6903	3	negative	5509	0	negative	10770	0	negative
	Repeat 2	21673	241	1.0998%	6757	2	negative	5565	0	negative	9911	0	negative
3	Repeat 1	23017	0	negative	4651	656	12.3610%	4622	0	negative	8788	0	negative
	Repeat 2	23485	0	negative	5066	692	12.0181%	5274	0	negative	9391	0	negative
4	Repeat 1	31688	0	negative	7203	0	negative	1032	996	49.1124%	13032	0	negative
	Repeat 2	30261	0	negative	8300	1	negative	1047	991	48.6261%	13937	0	negative

Correct positive calls are in ***bold italic***. Theoretical percentages are indicated in brackets.

269

270 Discussion

271 High-throughput sequencing is greatly enhancing the capacity to generate inexpensive
 272 and reliable genomic information. Illumina's bridge PCR chemistry is the most widely
 273 used clustering mechanism in high-throughput NGS. Illumina recently changed to
 274 ExAmp chemistry, which allows cluster generation to occur simultaneously with DNA
 275 seeding onto patterned arrays to minimize the likelihood that multiple library fragments
 276 are amplified in the same cluster. However, free adapters cannot be completely
 277 removed through purification, and with the presence of polymerase and templates,

278 index hopping can be initiated using false adapters[4] (**Fig. 1b**). Thus, sequencing
279 platforms utilizing ExAmp chemistry are at higher risk of index swapping between
280 samples in a multiplex pool[3, 4, 6]. A recent publication reports dramatically varied
281 index hopping rates with different library construction methods and also indicates that
282 these rates depend on machine types and flow cell batches[5]. PCR-free WGS had the
283 highest total contamination rate of ~6%[5]. Extra library clean-up, stringent filters, and
284 unique dual indexed adapters have been used to mitigate this problem[11, 14, 15].
285 Unique dual indexing moves more mis-assigned reads to the “filtered-out reads”
286 compared with regular combinatorial dual indexing. However, the empirical data from
287 Costello M. et al. demonstrated that double index switching could not be filtered out
288 efficiently even with unique dual indexing, and caused 1 error in 1223 reads[5]. Thus, in
289 spite of using unique dual indexes, the applications requiring high sensitivity for low
290 frequency allele detection or single cell sequencing would still be affected by the ExAmp
291 chemistry. Furthermore, this unique dual indexing approach requires complicated and
292 costly adapter and index design, more sequencing directions, and consequently
293 increased sequencing time and cost, and it limits the scalability of multiplexing large
294 numbers of samples.

295 However, not all sequencing platforms suffer from the index swapping issue. The
296 unique DNB technology used on BGI sequencers for making DNA copies is a linear
297 RCR amplification that is not prone to physical index hopping during DNB preparation
298 and arraying. There are two findings supporting this assertion. First, the empty controls
299 in the control test (index 33-40, Table 1) and in the HPV panel (index 8) have
300 exceptionally low index switching rates from one in 36 million (with filtering) to one in 5

301 million (without filtering). Second, in the WGS-like library preparation method, balancing
302 libraries with indexes 41-48 were mixed into the pooled libraries (index 1-8). Unlike the
303 mis-assignment of indexes 1-8, which includes all the contamination starting from library
304 preparation, the mis-assignment of indexes 41-48 only represents the steps after DNB
305 preparation. The average per-index mis-assignment rate for indexes 41-48 (Table 1) is
306 1 in 500,000 reads to 1 in 1,000,000 depending on quality filters, suggesting minimal
307 index mis-assignment during and after DNB preparation and arraying.

308 We have examined various protocols in detail and found that when pooling is
309 performed after PCR amplification, the index split rates are highly uniform; both index
310 cross-talk in empty controls and total mis-assignment rates are extremely low.
311 Removing apparent oligo synthesis errors can further reduce the total mis-assigned
312 reads by 32%, indicating that oligo quality is most likely the major cause of the
313 remaining index mis-assignment on BGI sequencers. Because single indexing would be
314 affected by oligo quality to a greater extent compared with unique dual indexing, high-
315 quality oligo without any contamination or errors (e.g., nucleotide deletions) is required
316 for the detection of ultralow levels of DNA or diagnostic DNA in DNB-based NGS
317 platforms.

318 We propose the following practices to maximally avoid index contamination: 1)
319 order TruGrade-equivalent ultrapure oligos to minimize contamination or artifacts and
320 validate the indexes using an NGS QC method if possible; 2) pool libraries after PCR
321 amplification; 3) apply a Q30 filter to increase accuracy by removing most sequencing
322 errors, although the quantity of total reads may decrease. Using this strategy, the actual
323 individual index mis-assignment rate on the BGI sequencing platform is only ~0.0001-

324 0.0004% with single indexing; this provides order(s) of magnitude higher precision
325 compared with the unique dual indexing method on newer Illumina platforms(12) and it
326 involves a much simpler adapter structure and fewer sequencing directions.

327 In summary, the DNB-based NGS platform has rare background-level single
328 index mis-assignment in all frequently used library construction methods we tested,
329 including WGS-like with PCR, PCR-free WGS-like, and two-step targeted PCR libraries,
330 ensuring the best data quality for the NGS community. Single DNB indexing provides a
331 simple and economical solution for large scale multiplexing, thus aiding more efficient
332 clinical research.

333

334 **Methods**

335 **WGS-like NGS Library Preparation**

336 Approximately 400-bp fragments of eight genes (**Fig. 2b** and Supplemental
337 **Table 1**) were individually amplified by rTaq (Takara Bio, Inc.) and size selected with a
338 2% agarose gel (Bio-Rad). Following Agencourt AmpureXP bead purification and
339 quantification with the Qubit™ dsDNA HS Assay kit (Thermo Fisher Scientific), single 3'-
340 A overhangs were added to 100 ng of PCR products through an in-house dA-tailing
341 reaction at 37°C for 30 minutes; heat inactivation was then performed at 65°C for 15
342 min. Adapter ligation was performed at 25°C for 30 minutes in a proprietary ligation
343 mixture containing 1.25 μM indexed adapters (regular oligo synthesis through IDT). In
344 the control test, eight empty controls individually tagged with indexes 33 to 40 were

345 incubated with water instead of PCR products for ligation. For Ad-1B- and Ad-2B-pooled
346 libraries, equal masses of the ligated samples with indexes 1 to 8 were mixed after one
347 or two rounds of bead purification, respectively. For all libraries, whether pooled or not,
348 PCR was performed using 1x KAPA HIFI Hotstart ReadyMix (KAPA) and PCR primers
349 (Invitrogen). After 5 cycles of amplification, 80 µL of beads was added to 100 µL PCR
350 reactions to clean the reaction. Samples of 20 ng of PCR products with individual
351 indexes were then mixed and used as PCR-pooled libraries. A total of 160 ng of PCR
352 products was used to form single strand circles (ssCir), 10 ng of which was used to
353 prepare DNBs using the SOPs for BGISEQ-500(8). We also pooled indexed samples at
354 equal quantities after ssCir formation (ssCir-pooled libraries) and after DNB preparation
355 (DNB-pooled libraries) based on Qubit™ ssDNA quantification. To balance the
356 positional base compositions for sequencing needs, 10 ng of ssCir from a human WGS
357 library control with indexes 41-48 (Invitrogen, China) was added to the ssCirs of Ad-,
358 PCR- or ssCir-pooled libraries. DNB-pooled libraries were mixed with the balancing
359 library immediately after DNB preparation. This balancing WGS library was constructed
360 as reported previously(8). Each pooling strategy was repeated in triplicate and
361 sequenced for single-end reads of 30 bp and index reads of 10 bp on the BGISEQ-500
362 platform.

363 **HPV Library preparation**

364 Control plasmid DNA containing individual HPV genotype 11, 18, 31, 33, 45, or
365 52 or combinations of these was diluted to 1,000 copies per sample and mixed with 5
366 ng of YH genomic DNA (**Table 2a, Supplementary Table 5**). These positive control

367 samples were used in three triplicate experiments. YH genomic DNA alone was used as
368 an HPV-negative control, and water was used as a multiplex PCR negative control.
369 Each sample was amplified and tagged individually with a 10-bp MGI sample index
370 during PCR using the BGI SeqHPV panel, which recognizes a broad spectrum of HPV
371 genotypes and β -globin derived from the *HBB* gene. Multiplex PCR was performed in a
372 96-well plate (Axygen). Twelve amplified samples were pooled into one, and then bead
373 purification was performed. The amplified DNA was provided with a 3'-A overhang and
374 ligated to a dT-tailed adapter containing index 1 to 6 independently as described above.
375 Empty controls with water were ligated with adapters containing index 7 or 8. After ssCir
376 formation, DNA with indexes 1 to 8 was pooled using equal volumes and purified after
377 digestion with exonucleases. The ssCir of the balancing library with indexes 41 to 48
378 was again added to the ssCirs of pooled experimental samples. The triplicates were
379 sequenced using 100 bp + 10 bp single-end runs on BGISEQ-500.

380 **Cancer Panel Library Preparation**

381 Reference standard DNA amplified from three NSCLC cell lines was purchased
382 from Horizon Diagnostics (Cambridge, UK), including the following: EGFR L858R (Cat.
383 ID: HD254), KRAS G12D (Cat. ID: HD272), and EGFR Δ E746-A750 (Cat. ID: HD251).
384 The DNA carrying EGFR L858R, KRAS G12D, or EGFR Δ E746-A750 mutations was
385 spiked into wild-type YH genomic DNA at ratios of 1%, 10%, or 50%, respectively. YH
386 genomic DNA alone was included as a negative control. A proprietary two-step PCR
387 protocol was used to enrich 181 COSMIC variant loci covered by MGI's lung cancer
388 panel kit (BGI). During thermal cycling, a sample index and molecular UIDs were

389 introduced to individual targeted regions. The indexed oligos used in this assay were
390 purchased from IDT through the TruGrade service. The purified multiplex PCR products
391 were validated on a Qubit fluorometer (Thermo Fisher), pooled with equal mass, and
392 used to prepare ssCirs and DNBs using standard procedures. A balancing WGS control
393 library was mixed after ssCir formation. The duplicated libraries were sequenced for
394 paired-end 50-bp reads along with a 10-bp index region.

395 **Sample QC and NGS statistics**

396 Raw data in FASTQ format obtained from BGISEQ-500 were split into separate
397 FASTQ files based on specific indexes with 0 bp (for control test) or 1 bp (for all other
398 WGS tests) of allowed mismatch. After FASTQ files with individual indexes were
399 generated, the third BWA algorithm, bwa aln, was then used to align the reads to the
400 human reference genome *hg38*. BAM files from bwa alignment were analyzed to
401 calculate the contamination rates. The reads with proper combinations of index and
402 amplicon were counted and highlighted in green in Fig. 3c. The reads mismatched to
403 incorrect genomic regions were collected for further error type analysis. The base score
404 Q30 (Sanger Phred+33 quality score) was used to assess the sequencing quality at
405 both genomic and index regions. By applying different Q30 filters to the index
406 sequences, we managed to reduce the number of reads with sequencing errors by at
407 least two-fold, and more than 96% of total reads remain with high quality (**Fig. 2b and**
408 **Supplementary Table 3**). Total index contamination equals the sum of all hopped
409 reads (data with brown shading) divided by the total reads of all the indexes shown in
410 the tables.

411 For HPV tests, the raw data were preprocessed based on information from lanes
412 and adapters. Using perfectly matched index reads, fq.gz raw sequencing reads were
413 then re-assigned to each sample, and at the same time index and primer sequences
414 were removed. The remaining reads from targeted PCR were aligned to the reference
415 sequences of *HBB* and various HPV types using bwa aln. Matched reads no fewer than
416 the corresponding cut-off were called positive.

417 In the cancer panel, raw FASTQ reads were analyzed by SOAPnuke (version
418 1.5.6). After trimming the adapter and removing low-quality reads, unique identifier
419 sequence information was retrieved and added into the sequence ID of the clean
420 FASTQ data by an in-house developed bioinformatic pipeline. We also calculated the
421 mapping rate, capture rate (fraction of target reads in all reads), duplication rate, and
422 uniformity (fraction of the amplicons whose depth exceeds 20% of the average depth in
423 all amplicons). After removing duplication, a BAM file was generated; variant calling was
424 performed by in-house developed software, and indel calling was performed using
425 Genome Analysis Toolkit (v4.0.3.0, GATK Mutect2).

426

427 **Abbreviations**

428 **WGS**: whole genome sequencing **NGS**: next generation sequencing
429 **DNB**: DNA-nanoball **cPAS**: combinatorial Probe Anchor Synthesis
430 **RCR**: rolling circle replication **ssCir**: single-stranded circle
431 **UID**: unique identifier **QC**: quality control **SD**: standard deviation

432

433 **Declarations**

434 **Acknowledgments**

435 We would like to acknowledge the ongoing contributions and support of all Complete
436 Genomics and BGI-Shenzhen employees, in particular the many highly skilled
437 individuals that build the BGI sequencers and work in the libraries, reagents, and
438 sequencing groups and make it possible to generate high-quality whole genome data.

439 **Funding**

440 This work was supported in part by Shenzhen Peacock Plan
441 No.KQTD20150330171505310.

442 **Availability of data and materials**

443 The dataset supporting the conclusions of this article is available at EBI-ENA, under
444 accession ID (CNSA): CNP0000071 and accession ID (ENA): PRJEB27504.
445 All the other data used here are included within the published article and its Additional
446 files.

447 **Authors' contributions**

448 QL, XZ, WZ and YJ designed experiments of the study. QL, XZ and HS performed
449 experiments. LW prepared the tables, figures and drafted the manuscript, YJ supervised
450 the project and manuscript editing. DX, ZM, QL, SD, ZL assisted with bioinformatic
451 analysis. All authors read and approved the final manuscript.

452 **Ethics approval**

453 NO. BGI-R027

454

455 **Competing interests**

456 Employees of BGI and Complete Genomics have stock holdings in BGI.

457

458 **References**

459 1. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-**
460 **generation sequencing technologies.** *Nat Rev Genet* 2016, **17**(6):333-351.

461 2. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon
462 DR, Ordoukhalian P: **Library construction for next-generation sequencing:**
463 **overviews and challenges.** *Biotechniques* 2014, **56**(2):61-64, 66, 68, *passim*.

464 3. **Effects of Index Misassignment on Multiplexing and Downstream Analysis**
465 **(white paper)** [<https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf>]

468 4. Rahul Sinha GS, Gunsagar Singh Gulati, Camille Ezran, Kyle Joseph Travaglini,
469 Eric Wei, Charles Kwok Fai Chan, Ahmad N Nabhan, Tianying Su, Rachel Marie
470 Morganti, Stephanie Diana Conley, Hassan Chaib, Kristy Red-Horse, Michael T
471 Longaker, Michael P Snyder, Mark A Krasnow, Irving L Weissman: **Index**
472 **Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In**
473 **Illumina HiSeq 4000 DNA Sequencing.** In. Edited by Medicine SUSo. bioRxiv;
474 2017.

475 5. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B,
476 Green L, Howd T, Mason T *et al*: **Characterization and remediation of sample**
477 **index swaps by non-redundant dual indexing on massively parallel**
478 **sequencing platforms.** *BMC Genomics* 2018, **19**(1):332.

479 6. Vodak D, Lorenz S, Nakken S, Aasheim LB, Holte H, Bai B, Myklebost O, Meza-
480 Zepeda LA, Hovig E: **Sample-Index Misassignment Impacts Tumour Exome**
481 **Sequencing.** *Sci Rep* 2018, **8**(1):5307.

482 7. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, Qu S, Mei X, Chen H, Yu T *et al*:
483 **A reference human genome dataset of the BGISEQ-500 sequencer.**
484 *Gigascience* 2017, **6**(5):1-9.

485 8. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML:
486 **Genome-wide genetic marker discovery and genotyping using next-**
487 **generation sequencing.** *Nat Rev Genet* 2011, **12**(7):499-510.

488 9. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E,
489 Shendure J, Turner DJ: **Target-enrichment strategies for next-generation**
490 **sequencing.** *Nat Methods* 2010, **7**(2):111-118.

491 10. Andolfatto P, Davison D, Erez Yilmaz D, Hu TT, Mast J, Sunayama-Morita T,
492 Stern DL: **Multiplexed shotgun genotyping for rapid and efficient genetic**
493 **mapping.** *Genome Res* 2011, **21**(4):610-617.

494 11. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, Light M,
495 Lai K, Jarosz M, McNeill MS *et al*: **Unique, dual-indexed sequencing adapters**
496 **with UMIs effectively eliminate index cross-talk and significantly improve**
497 **sensitivity of massively parallel sequencing.** *BMC Genomics* 2018, **19**(1):30.

498 12. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: **Detection and**
499 **quantification of rare mutations with massively parallel sequencing.** *Proc*
500 *Natl Acad Sci U S A* 2011, **108**(23):9530-9535.

501 13. Christensen E, Nordentoft I, Vang S, Birkenkamp-Demtroder K, Jensen JB,
502 Agerbaek M, Pedersen JS, Dyrskjot L: **Optimized targeted sequencing of cell-**
503 **free plasma DNA from bladder cancer patients.** *Sci Rep* 2018, **8**(1):1917.

504 14. Wright ES, Vetsigian KH: **Quality filtering of Illumina index reads mitigates**
505 **sample cross-talk.** *BMC Genomics* 2016, **17**(1):876.

506 15. Kircher M, Sawyer S, Meyer M: **Double indexing overcomes inaccuracies in**
507 **multiplex sequencing on the Illumina platform.** *Nucleic Acids Res* 2012,
508 **40**(1):e3.

509

510

511 **Figure legends**

512 **Figure 1: Mechanisms of index hopping on different sequencing platforms. (a)**
513 Sequencing using DNA nanoball technology is accomplished through Phi29 and RCR
514 linear amplification; each copy is amplified independently using the same template
515 ssCir. In this case, error reads from index hopping cannot accumulate, and most of the
516 signal originates from correct indexes. (b) Bridge PCR or ExAmp chemistry utilizes
517 exponential amplification, and index hopping can accumulate as amplification proceeds
518 through each cycle, resulting in mis-assigned samples. Green, correct index; red, wrong
519 index.

520

521 **Figure 2: Library preparation workflows.** (a) “standard PCR-based WGS”-like library;
522 (b) PCR-free library; (c) two-step PCR library. Pooling after each step, indicated by red
523 arrows, is examined for different library preparation strategies. Gray rectangle, adapter;
524 colored rectangle, unique index assigned to a particular sample; gray vertical lines,
525 unique sample index; white rectangle, UID.

526

527 **Figure 3: a. Total contamination rates for each pooling scenario.** Three replicates
528 are presented with different types of bars. Wider bars with dashed borders represent the
529 average of the three replicates, the exact values of which are labeled on top. **b. Index**
530 **split rates when pooling was performed after PCR amplification.** Average \pm
531 standard deviation (SD) of three replicates is presented. The theoretical split rate for
532 each index is 0.125. **c. Index contamination matrix when pooling occurred after**
533 **PCR purification.** Indexes 1 to 8 were assigned to Notch1, EFEMP2, Lox, USP9Y,
534 HIST1H1D, C7orf61, GXYLT2, and TM9SF4 respectively. Read numbers and
535 percentages are shown with or without Q30 filter application. Green shading, proper
536 combinations; brown and yellow shading, improper combinations; yellow shading,
537 improper combinations likely resulting from contamination during oligo synthesis. Index
538 contamination rates were calculated by dividing the sum of contaminated reads by the
539 sum of total reads for all eight indexes.

540

541 **Figure 4: The effect of filter on total contamination rate and percent of remaining**
542 **reads.** The reads when library pooling occurred after PCR amplification were filtered.

543 Total contamination rate is shown in red and percent of remaining reads is shown in
544 blue. Reads with index 7 were excluded from the calculation. Mapped reads were
545 filtered by different criteria for the Q30 score. Averages \pm SD of three replicates are
546 presented. The average values are labeled on top.

547

548 **Supplementary information**

549 Supplementary Table 1. PCR primer sequences for 8 genes.

550 Supplementary Table 2. Total reads and rates of all WGS libraries (indexes 1-8).

551 Supplementary Table 3. Effect of Q30 filter on sequencing reads and rates when library
552 pooling is performed after PCR amplification (indexes 1-8).

553 Supplementary Table 4. Index contamination in water control with PCR-free library.

554 Supplementary Table 5. Raw data of PCR-free library contamination, 3 lanes.

555

556

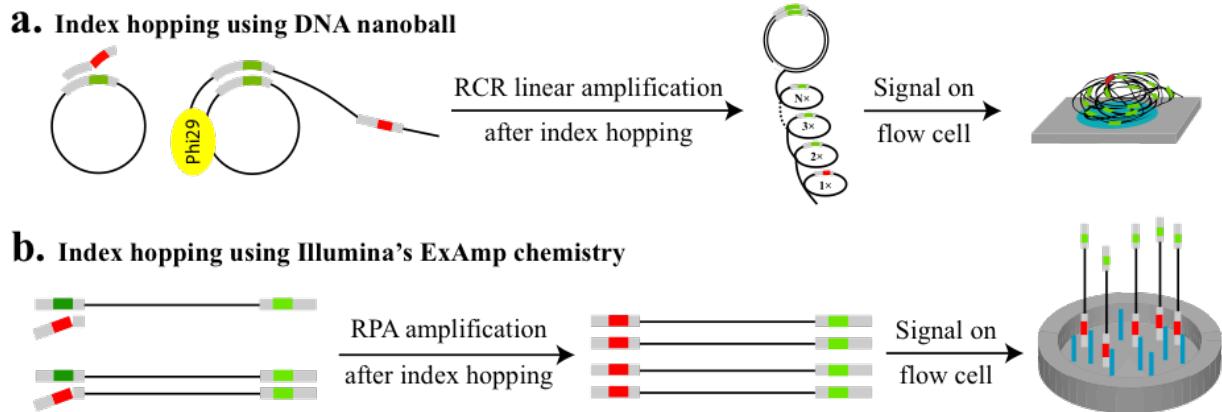


Figure 1

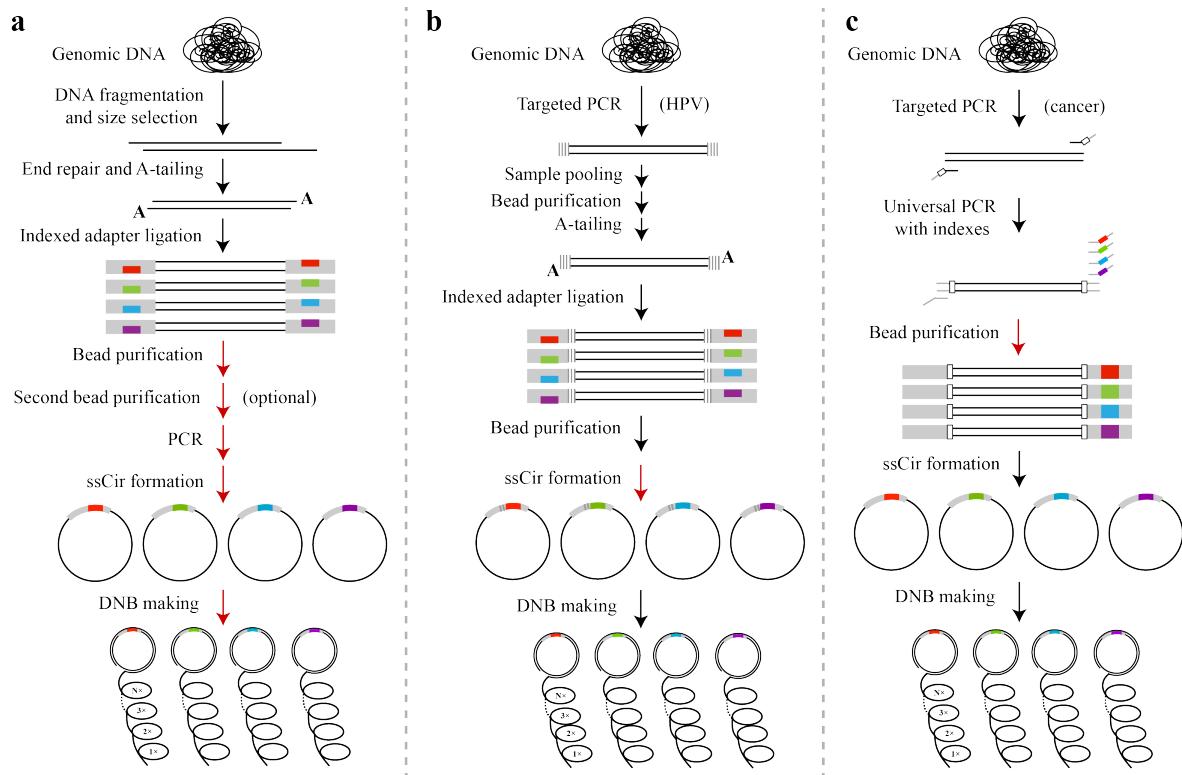


Figure 2

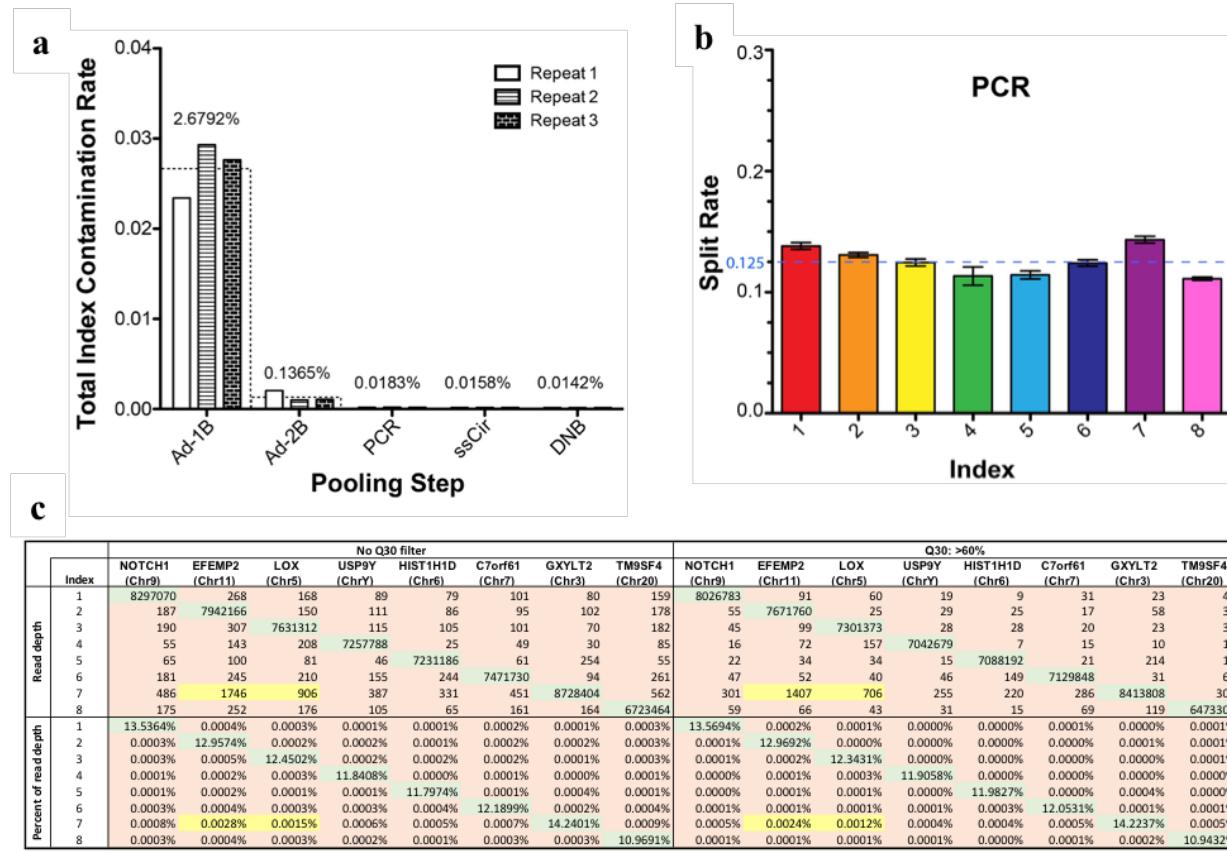


Figure 3

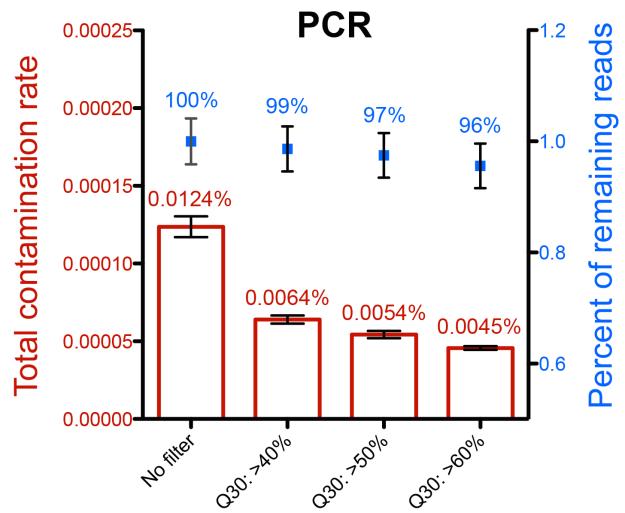


Figure 4