Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies

Tom van der Valk[1], Francesco Vezzi[2], Mattias Ormestad[2], Love Dalén[3], Katerina Guschanski[1]**

[1]Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden

[2]Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden

[3]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, SE-10405 Stockholm, Sweden

* These authors contributed equally

Corresponding authors: Katerina Guschanski (katerina.guschanski@ebc.uu.se), Tom van der Valk (tom.vandervalk@ebc.uu.se)

Abstract

The high-throughput capacities of the Illumina sequencing platforms and the possibility to label samples individually have encouraged a wide use of sample multiplexing. However, this practice results in read misassignment (usually <1%) across samples sequenced on the same lane. Alarmingly high rates of read misassignment of up to 10% were reported for the latest generation of Illumina sequencing machines. This may make future use of the newest generation of platforms prohibitive, particularly in studies that rely on low quantity and quality samples, such as historical and archaeological specimens. Here, we rely on barcodes, short sequences that are ligated to both ends of the DNA insert, to directly quantify the rate of index hopping in 100-year old museum-preserved gorilla (*Gorilla beringei*) samples. Correcting for multiple sources of noise, we identify on average 0.470% of reads containing a hopped index. We show that sample-specific quantity of misassigned reads depend on the number of reads that any given sample contributes to the total sequencing

pool, so that samples with few sequenced reads receive the greatest proportion of misassigned reads. Ancient DNA samples are particularly affected, since they often differ widely in endogenous content. Through extensive simulations we show that even low index-hopping rates lead to biases in ancient DNA studies when multiplexing samples with different quantities of input material.

Keywords

Read misassignment, next generation sequencing, ExAmp chemistry, multiplexing, museum specimens, index switching

31 **1 | Introduction**

32 Multiplexing samples for sequencing is common practice in genomic studies (Craig *et al.* 2008; Meyer
33 & Kircher 2010; Smith *et al.* 2010). During multiplexing, samples are individually labelled with unique
34 identifiers (indices) that are embedded within one (single indexing) or both (dual indexing)
35 sequencing platform-specific adapters (Meyer & Kircher 2010; Kircher *et al.* 2012). The samples are
36 subsequently pooled into a single DNA library and sequenced on the same lane, greatly reducing per
37 sample sequencing cost. Following sequencing, computational demultiplexing based on the sample-
38 specific indices enables the assignment of sequenced reads to the respective sample of origin. In
39 recent years, the output from sequencing platforms has increased dramatically, making multiplexing
40 the recommended standard sequencing workflow on the latest generation of Illumina platforms (e. g.
41 NovaSeq) (Illumina Inc.). However, ever since multiplexing approaches were introduced, low rates of
42 read misassignment across samples sequenced on the same lane have been reported on all Illumina
43 platforms (Kircher *et al.* 2012; Nelson *et al.* 2014; Renaud *et al.* 2015; Wright & Vetsigian 2016;
44 D'Amore *et al.* 2016). Read misassignment is the result of reads carrying an unintended index and
45 consequently being erroneously attributed to the wrong sample. Processes resulting in read
46 misassignment, i.e. presence of reads with an incorrect index, are numerous. The effect of
47 sequencing errors that can convert one index sequence into another is well known and has led to
48 series of recommendations for designing highly distinct indices (Meyer & Kircher 2010). Jumping PCR
49 during bulk amplification of library molecules that carry different indices can generate chimeric
50 sequences and should be avoided (Meyerhans *et al.* 1990; Odelberg *et al.* 1995; Carlsen *et al.* 2012;
51 Esling *et al.* 2015). Similarly, cross-contamination of indexing adapters during oligonucleotide
52 synthesis or laboratory work can lead to reads obtaining an unintended index. Additionally, cluster
53 misidentification due to "bleeding" of indices into neighbouring clusters have been reported on all
54 high throughput sequencing platforms (Kircher *et al.* 2012; Nelson *et al.* 2014; Renaud *et al.* 2015;
55 Mitra *et al.* 2015; Wright & Vetsigian 2016; D'Amore *et al.* 2016; Vodák *et al.* 2018). For the latest
56 Illumina platforms with patterned flow cells and ExAmp chemistry (e.g. Hiseq X and NovaSeq), it has

57 been suggested that read misassignment is caused by the presence of free-floating indexing primers
58 in the final sequencing library (Illumina Inc. 2017; Sinha *et al.* 2017). Such free-floating molecules can
59 appear if sequencing libraries are not stored properly and become fragmented, or if the final
60 sequencing libraries retain non-ligated indexing primers due to inefficient clean-up and size selection
61 (Illumina Inc. 2017). These free-floating primers can anneal to the pooled library molecules and
62 become extended by the DNA polymerase before the rapid exclusion amplification on the flow cell,
63 creating a new library molecule with an erroneous index (Figure 1). We refer to this particular
64 process of generating misassigned reads as **index hopping**. The reported rate of read misassignment
65 on Illumina platforms that rely on the traditional bridge amplification for cluster generation is low
66 (<1%) (Kircher *et al.* 2012; Nelson *et al.* 2014; Wright & Vetsigian 2016) and therefore this source of
67 error has been readily ignored. However, on the latest Illumina patterned flow cell platforms with
68 ExAmp chemistry, the reported rate of read misassignment ranges from 0% to 10% (Illumina Inc.
69 2017; Sinha *et al.* 2017; Owens *et al.* 2018; Griffiths *et al.* 2018; Vodák *et al.* 2018), with Illumina
70 quoting a read misassignment rate of up to 2% (Illumina Inc. 2017).

71 As a consequence of conflicting results, the prevalence and severity of read misassignment on the
72 latest Illumina platforms remain unclear. This is partly due to the difficulties of reliably identifying
73 misassigned reads in sequencing experiments, particularly when pooling similar samples types (e.g.
74 multiple individuals from the same population that have high sequence similarity). The use of dual
75 indexing allows for the filtering of the majority of reads that show signs of read misassignment
76 (Kircher *et al.* 2012). However, since indices can potentially be switched at both ends of the molecule
77 and the number of available indices is limited, it remains difficult to directly quantify read
78 misassignment rates on these platforms. Consequently, the so far reported rates off index-hopping
79 have been estimated using indirect methods (Sinha *et al.* 2017; Owens *et al.* 2018; Griffiths *et al.*
80 2018; Vodák *et al.* 2018).

81 The reported high rates of index-hopping are especially worrisome for studies involving sequencing
82 data obtained from degraded samples such as ancient and historical specimens, since in most cases
83 such studies rely on low-coverage genomic data (Shapiro & Hofreiter 2014). Inferences are therefore
84 often based on subtle differences between limited sets of polymorphic sites, so that even small
85 quantities of misassigned sequencing reads can potentially lead to erroneous conclusions. It is thus
86 crucial to distinguish genuine sample-derived endogenous DNA fragments from false signals
87 (Skoglund *et al.* 2014).

88 The purpose of this study is two-fold. First, we aim to directly quantify the rate of index-hopping on
89 the Illumina patterned flow cell platforms for a standard ancient DNA library. To this end, we make
90 use of inline barcodes, short unique seven base pair sequences ligated to both ends of the DNA
91 fragments (Rohland & Reich 2012), in combination with the indexed primers that are traditionally
92 used for sample identification. The barcodes become part of the sequencing read and thus allow for
93 accurate identification of the read origin, even in the presence of index hopping. Therefore, the
94 amount of index hopping can be directly quantified by identifying reads with wrong barcode-index
95 combinations. Second, we aim to identify and characterize biases resulting from index-hopping in
96 pooled ancient DNA libraries that may impact downstream analyses typical for ancient DNA research.
97 To achieve this, we simulate ancient DNA sequencing libraries under different rates of index-hopping
98 and quantify the impact of misassigned reads on population genomic inferences by performing a set
99 of standard genome-wide analyses.

100 **2| Methods**

101 **2.1| Library preparation**

102 DNA extracts from seven historical eastern gorilla (*Gorilla beringei*) samples were turned into
103 sequencing libraries as described in (van der Valk *et al.* 2018) (see supplementary material). All
104 library preparation steps except indexing PCR were performed in a dedicated ancient DNA laboratory
105 to minimize contamination. Briefly, 20 µl DNA extract was used in a 50 µl blunting reaction together
106 with USER enzyme treatment to remove uracil bases resulting from aDNA damage (Briggs *et al.*
107 2010). DNA fragments within each sample were then ligated to a unique combination of incomplete,
108 partially double-stranded P5- and P7-adapters, each containing a unique seven base pair sequence
109 (Rohland *et al.* 2015) (Table S1). We refer to these as the P5 and P7 **barcodes** from here on. All
110 barcode sequences were at least three nucleotides apart from each other to ensure high certainty
111 during demultiplexing and avoid converting one barcode into another through sequencing errors
112 (Rohland *et al.* 2015) (Table S1). To increase the complexity of the pooled sequencing library, one
113 sample (sample 7) was split in two fractions, each of which received a different barcode combination
114 (Table 1).

115 Indexing PCR was performed for 10 cycles using a unique P7 indexing primer for each sample, as in
116 (Meyer & Kircher 2010) (Table S1). We refer to the unique sequence added during the indexing PCR
117 as the P7 **index**. As with the barcodes, all index sequences differed by at least three base pairs from
118 each other (Table S1). Indexing PCR for sample 7 was performed in a single reaction combining both
119 fractions of this sample. Following the indexing PCR, each DNA fragment contained three unique
120 identifiers: the P5 and P7 barcodes directly ligated to the ends of the DNA fragments, and the P7
121 index contained within the Illumina sequencing adapter (Figure 1). Sample libraries were cleaned
122 using MinElute spin columns, fragment length distribution and concentrations were measured on the
123 Bioanalyzer. We then pooled all seven sample libraries in a ratio of 2:1:2:1:1:1:2 for samples 1 to 7,
124 and performed two rounds of AMPure XP bead clean-up, using 0.5X and 1.8X bead:DNA ratio,

125 respectively. We confirmed that indexing primers were successfully removed during clean-up by
126 running the final library on a Bioanalyzer (Figure S1). The pooled library was sequenced on three
127 HiSeqX lanes that were part of independent runs with a 5% phiX spike-in, at the Science for Life
128 Laboratory in Stockholm.

129 **2.2| Data processing**

130 All reads were demultiplexed based on their unique indices using Illumina's bcl2fastq (v2.17.1)
131 software with defaults settings, allowing for one mismatch per index and only retaining "pass filter"
132 reads (Illumina Inc.). All unidentified reads, i.e. reads containing indices not used in our experiment,
133 were retained and subjected to the same filtering steps as assigned reads (see below). We removed
134 adapter sequences using AdapterRemoval V2.1.7 with standard parameters (Schubert *et al.* 2016).
135 Due to the fragmented nature of DNA in historical samples, we could subsequently merge the reads,
136 requiring a minimal overlap of 11bp and allowing for a 10% error rate. The merging of reads allowed
137 us to obtain sequencing information for the complete DNA molecule and thus to accurately identify
138 the barcodes on both ends of the DNA fragment (P5 and P7 barcodes, respectively, Figure 1).
139 Unmerged reads and reads shorter than 29 basepairs were removed. To increase certainty, we only
140 retained reads with error-free P5 and P7 barcodes and an average quality score of at least 30 using
141 prinseq V0.20.4 (Schmieder & Edwards 2011).

142 **2.3| Disentangling cross-contamination from index hopping**

143 Low rates of cross-contamination of barcodes and indexes can be expected, even if strict measure
144 are followed during library preparation, such as the use of clean-room facilities (Kircher *et al.* 2012).
145 This can result in reads containing a wrong index-barcode pair and could be falsely interpreted as
146 evidence for index hopping. Since the inline barcodes used in this experiment are unaffected by
147 index-hopping (Figure 1), we can accurately estimate the rate of barcode cross-contamination as the
148 fraction of reads containing a P5-P7 barcode pair that was not used during library preparation. In rare
149 cases, barcode cross-contamination results in a read with a valid barcode pair (e.g. a barcode

150 combination that was intendedly used during library preparation) and thereby remain undetected in
151 our estimate. However, since we used every barcode only once, the proportion of reads resulting
152 from such an event is several orders of magnitude lower than the fraction of reads containing an
153 invalid barcode pair and does therefore not significantly affect any of our estimates (see
154 supplementary material).

155 As the Illumina HiSeq X platform did not support a dual-indexing design at the time of this
156 experiment, the rate of index cross-contamination could not be estimated using invalid index pairs.
157 Therefore, we relied on the fact that of the 40 indices that are routinely used in our laboratory only
158 seven were implemented in this experiment (Table S2). Assuming a relatively equal rate of cross-
159 contamination between all 40 indexes, we estimated index cross-contamination as the fraction of
160 reads containing any of the 33 indices that were not deliberately included during our experiment.

161 We then determined the raw rate of index hopping as the fraction of reads showing an index-
162 barcode combination not used during the library preparation. We accounted for the possibility of
163 barcode and index cross-contamination resulting in the same barcode-index combination by
164 subtracting the contamination estimates obtained above from the raw value of index hopping. All
165 statistical analyses were performed in R 2.15.3 (Team R Core 2016) (see supplementary material).

166 **2.4| Simulations of aDNA sequence libraries**

167 To quantify downstream biases resulting from index-hopping during pooled sequencing, we
168 simulated ancient DNA sequencing libraries with different endogenous content under varying rates
169 of index hopping. First, four "template" genomes were simulated to serve as seeds for four
170 populations, popA, popB, popC and popD, respectively, by using chromosome 1 of the gorilla
171 reference (removing all N nucleotides) (Gordon *et al.* 2016). The population divergence was set as
172 follows: popA – popB: 20.000 years, popC – popD: 20.000 years and popA/popB – popC/popD:
173 200.000 years (Figure 3b) and we introduced random mutations at a rate of $1.67 \cdot 10^{-9}$ per base per
174 year (corresponding to the estimated gorilla mutation rate (Besenbacher *et al.* 2018)). We then

175 simulated thirty individuals for each population, using the "template" genomes as a starting point
176 and introducing on average $5 \cdot 10^{-7}$ random mutations per base (corresponding to all individuals
177 within each population sharing a common ancestor on average 300 years ago). We did not simulate
178 any admixture between the populations. Next, each individual genome was converted into an
179 ancient DNA sequence library (fastq-format), with insert size normally distributed around 50bp and
180 endogenous content of either 0.133%, 0.398%, 1.33%, 3.98%, 13.26% or 39.78% to mimic
181 characteristics often observed in ancient DNA studies. The levels of endogenous content were
182 chosen to result in commonly observed genome coverages of ancient DNA samples (see below). The
183 non-endogenous reads consisted of fastq-reads simulated using the PhiX-reference genome (NCBI
184 nucleotide ID: NC_001422) as template. We then simulated sequencing output of equimolar pooled
185 libraries as would be obtained from sequencing the pools on four NovaSeq6000 runs (flow cell-type
186 S4, expected output 8-10 billion reads per run). The expected output of ~40 billion reads thus
187 consisted of a random sample of ~333 million fastq-reads from each of the 120 simulated sequencing
188 libraries (30 individuals x 4 populations). Index-hopping was simulated by giving each read a
189 predefined probability of randomly hopping into another sample, using the following rates: 0.0%,
190 0.1%, 0.5%, 1%, 5% and 10%, reflecting the levels of index-hopping reported in the literature. We did
191 not simulate indels/deletions, PCR duplications, sequencing errors, and post-mortem DNA damage,
192 since we specifically aimed to address the biases resulting from index-hopping. Our final simulated
193 data thus consisted of six datasets of 120 simulated ancient DNA libraries with varying endogenous
194 content obtained from four different populations, with a different level of index-hopping in each of
195 the six datasets.

196 To analyse the simulated data, we aligned all reads per individual to the gorilla reference
197 chromosome 1 (Gordon *et al.* 2016) (note that in our simulations this reference represents the
198 ancestral state for each site) using bwa-mem on default parameters (Li & Durbin 2009). The obtained
199 coverage for the simulated individuals was 0.1X, 0.3X, 1X, 3X, 10X or 30X, depending on the sample's
200 endogenous content (note that the levels of endogenous content were chosen to result in these

201 coverages). Next, we employed a pipeline specifically designed for analysing low-coverage genomes
202 from degraded DNA sources. We obtained genotype likelihoods for each individual using angsd
203 (Korneliussen *et al.* 2014), filtering reads below mapping quality of 30 (-minMapQ 30), a flag above
204 255 (-remove_bads 1) and removing reads with multiple hits (-uniqueOnly 1). We then only
205 considered genotypes with a likelihood ratio test statistic of minimum 24 (-SNP_pval 2e-6) using the
206 samtools genotype model (-GL 1).

207 **2.5| Inferring population genomic statistics from simulated data**

208 We used Principal Components Analysis, Admixture and D-stats (ABBA-BABA test) to reconstruct
209 population divergence under different levels of index-hopping. Principal Components Analysis was
210 run using PCAngsd with default parameters and 200 EM iterations for computing the population
211 allele frequencies (Fumagalli *et al.* 2013). Individual admixture proportion were obtained using
212 NgsAdmix (Skotte *et al.* 2013) at default parameters and using K = 4 (number of ancestral clusters).
213 Pairwise D-stats of the format (popA,popB,popC,ancesteral) were calculated for each possible pair of
214 individuals in popA and popB by sampling a random allele at each site (htsbox pileup -R -q 30 -Q 30 -s
215 1, https://github.com/lh3/htsbox), using a high coverage (30X) individual from popC as the third
216 ingroup and the ancestral allele (reference allele) at each site as the outgroup. Standard-deviations
217 and resulting Z-scores (the number of standard deviations of D from 0) were obtained using a
218 jackknife approach with blocksize of 2Mb.

219 **3 | Results**

220 **3.1 | Empirical data**

221 **3.1.1 | Barcode and index cross-contamination**

222 Since our sequencing libraries were made from degraded historical samples and thus contained a
223 large proportion of short DNA fragments (Figure S1) the majority of reads could be confidently
224 merged for all three sequencing runs (95.3% SE ± 1.0%). This allowed us to accurately infer both
225 barcodes at the read ends. After all filtering steps (Methods), the final dataset contained 89.3% SE ±
226 1.9% of the original sequence reads.

227 We estimate the average level of barcode cross-contamination across all three runs at 0.0276% SE ±
228 0.0026 (see methods, Table 1, Table S3, Figure S2), with different rates observed between samples
229 (global chi-square test, $P<10^{-15}$). Assuming that adapter ligation of barcodes is unbiased with respect
230 to the barcode sequence (Rohland *et al.* 2015), this low percentage of cross-contamination will lead
231 to a negligible fraction of reads ($1.09 \cdot 10^{-8}$%, see supplementary material) with a barcode pair that
232 wrongly appear as having undergone index hopping. The rate of index cross-contamination was
233 estimated at 0.124% SE ± 0.0023 (Table S4), by quantifying the fraction of reads containing indices
234 that were not intentionally used in our experiment (see Methods, Table S2).

235 **3.1.2 | The rate of index hopping**

236 Index hopping will not affect the barcodes that are directly ligated to the DNA fragments. Therefore,
237 it can be readily distinguished from barcode cross-contamination by identifying reads containing an
238 incorrect combination between an index and a barcode pair. Across all three sequencing runs, we
239 detected a low proportion of such reads (mean=0.594%, SE ± 0.0434%, Table 1). However, we
240 estimate that ~0.124% of these reads are a result of index and barcode cross-contamination (see
241 above). Therefore, the corrected rate of index hopping in our experiment across all three sequencing
242 runs is ~0.470% SE ± 0.044. The proportion of hopped reads differed significantly among samples

243 (chi-square test, $P<10^{-15}$) and was positively correlated with the number of sequenced reads per
244 sample (Pearson's r = 0.96, P = 0.0005, Figure 2). This suggests that in multiplexed sequencing runs,
245 the samples with higher number of sequenced reads will serve as the dominant source of hopped
246 reads. Even though the overall rate of index hopping is low, samples with proportionally few
247 sequenced reads are thus considerably more affected by index hopping. In our experiments, this
248 resulted in 2.49% SE ± 0.29% of hopped reads in the sample with the lowest number of sequenced
249 reads (Table S4, S5, Figure 2).

250 We find that the rate of index hopping differed significantly by read length and slightly by GC content
251 (chi-square test, both $P<10^{-15}$, Figure S3). Reads shorter than 90 bp and reads with GC content above
252 40% showed significantly higher proportion of hopped reads than expected under a random
253 distribution.

254 **3.2| Simulated data**

255 **3.2.1| Effects of index hopping on estimates of sample endogenous content**

256 Ancient DNA studies frequently rely on the screening of a large number of samples by means of
257 pooled low depth sequencing to identify samples with good DNA preservation and high endogenous
258 content. Through introduction of endogenous reads into low-quantity samples, index-hopping can
259 lead to a false signal of DNA preservation. We estimated the endogenous content for each of our
260 simulated ancient DNA sequencing libraries as the fraction of reads that mapped to the reference
261 genome under different rates of index-hopping (see Methods). We observed that already at low
262 rates of index hopping (<1%), the endogenous content of low-quality samples (0.1%-0.4%
263 endogenous content) was over-estimated (up to ~2-fold higher, Figure 3a). This bias became more
264 pronounced as rates of simulated index hopping increased and resulted in up to 8-fold higher
265 estimate of endogenous content. Estimates for samples with higher endogenous content (>3%) were
266 biased only at high rates of index-hopping (5-10%).

267 **3.2.2| Index hopping biases population genetic inferences**

268 Principal Components Analysis clearly differentiates the four simulated populations from each other
269 in the absence of index-hopping (Figure 3c, S4). However, we find that already at a low rate of index-
270 hopping (0.50%, similar to observed in our empirical data), population differences between popA-
271 popB and popC-popD start to disappear. This is caused by the relatively high number of hopped,
272 wrongly assigned reads in samples with low endogenous content. At extreme levels of index hopping
273 (10%), even the differences between the highly diverged populations (popA-popB vs popC-popD)
274 disappears (Figure 3c, S4).

275 Admixture analysis corroborated the results obtained from PCA. We find that at low rates of index-
276 hopping (0.50%), false signals of shared ancestry between individuals from different populations
277 start to appear in the low-coverage samples (Figure 3d, S5). At the highest rate of index-hopping
278 (10%), only the highest quality samples (e.g. 30X) remain unbiased (Figure 3d, S5).

279 We used ABBA-BABA counts to test if index hopping can lead to erroneous inferences of gene-flow
280 between the populations. Although the Z-scores become skewed at low rates of index-hopping (<1%)
281 if samples differ strongly in endogenous content, we only inferred significant deviations from zero at
282 high rates of index-hopping (>5%) (Figure S6).

4 | Discussion

Using a dual barcoding strategy during library preparation, we show that index hopping occurs on the Illumina HiSeq X platform, but its rate is low in our ancient DNA library (0.470% SE ± 0.044). Although multiple sources of error such as jumping PCR, barcode and index cross-contamination, sequencing errors, and index hopping can result in read misassignment, our experimental design allowed us to systematically address each of them. Jumping PCR can be eliminated as explanation for wrong index-barcode combinations, since we avoided amplification of pooled libraries from different samples. However, we show the strong effect of jumping PCR when looking at the rate of wrong barcode combinations in the only sample with two different barcode pairs that was amplified in a single indexing reaction (Fig. S2). We further show that the rate of barcode and index cross-contamination is very low (0.027% SE ± 0.0026 and 0.124% ± 0.0023, respectively) and therefore not the primary cause of observed reads with the wrong index-barcode pairs.

Read misassignment is not a novel phenomenon on the Illumina sequencing platforms. Reported error rates range from 0.1% to 0.582% on the HiSeq 2500 (Kircher *et al.* 2012; Wright & Vetsigian 2016) and from 0.06% to 0.51% on the MiSeq platforms (Nelson *et al.* 2014; Renaud *et al.* 2015; D'Amore *et al.* 2016). It is therefore noteworthy that the fraction of hopped reads as estimated in our study (0.470%) is similar to that reported for other platforms. However, it markedly differs from previous estimates for the Illumina platforms with ExAmp chemistry, which are based on sequencing modern (high quality) DNA and range from 0% to 2.5%-10% (Sinha *et al.* 2017; Owens *et al.* 2018; Griffiths *et al.* 2018). Since the sequencing chemistry of the Illumina NovaSeq platform is identical to that used for the HiSeq X, this platform is likely to be affected at a similar rate as reported here.

We used a standard library preparation protocol for degraded samples, which includes rigorous removal of free-floating adapters through size selection and cleaning (supplementary methods). This practice likely resulted in the relative low rate of index hopping in our experiment. As previously

307 suggested, strict library clean-up and size selection is thus recommended for multiplexed ancient
308 DNA sequencing studies.

309 A so far neglected observation is that the number of hopped reads into each sample is proportional
310 to the total number of reads contributed by this sample to the pooled sequencing library. Pooling
311 samples in different quantities leads to a greater proportion of hopped reads into samples with
312 fewer sequenced reads. In this study, libraries with the lowest number of sequenced reads displayed
313 up to 3.2% of misassigned reads (Table S5), an order of magnitude higher than the average rate
314 within a lane. The effect of this skewed rate of index hopping becomes even more severe if the
315 endogenous content is markedly different between samples. Since the endogenous content is usually
316 not known beforehand, pooling samples in equimolar quantities can lead to large differences in the
317 number of endogenous reads between samples. In such cases, even at the rate of index-hopping
318 reported here, the proportion of false assigned endogenous reads within low quantity samples can
319 reach rates above 10% (Figure S7), resulting in highly overestimated sample endogenous content
320 (Figure 3a). This is problematic, as presence of even few reads of interest can lead to further
321 processing and deep sequencing of an ancient DNA sample deemed to be of importance.
322 Additionally, we detected a higher rate of index hopping among shorter reads and small differences
323 in the fraction of index-hopped reads related to read GC-content. This suggests that the annealing of
324 free floating adapters present in the sequencing libraries does not occur randomly. The underlying
325 mechanisms are not yet well understood but could be related to differences in the DNA denaturation
326 temperatures between DNA fragments of different size. Due to the lower denaturation temperature,
327 short fragments might be occurring at a higher rate in single-stranded conformation and are thereby
328 more accessible to free floating index primers. Since shorter fragments in ancient DNA libraries often
329 represent endogenous DNA, whereas longer fragments are mostly environmental contamination
330 (Green *et al.* 2010), index-hopping can disproportionally affect the reads of interest in ancient DNA
331 libraries.

332 To further illustrate the effect of index hopping on estimates of endogenous content and population
333 genomics inferences, we employed simulations that encompass the complete range of reported
334 index hopping rates and span a distribution of endogenous content typical for ancient DNA studies.
335 Through simulations, we show that biases due to misassigned reads start to appear at index hopping
336 rates below 0.5% when analysing samples of low coverage (<3X). As samples with low endogenous
337 content predominantly act as receivers of hopped reads, inferences of population differentiation
338 become less clear (Figure 3c) due to many hopped reads being erroneously assigned to low quantity
339 samples. This also results in the false inference of shared ancestry between individuals from
340 divergent populations as exemplified by Admixture (Figure 3d). In contrast, the inference of gene
341 flow between populations through D-statistics is relatively robust to the biases resulting from index-
342 hopping, if the proportion of misassigned reads between the tested samples is similar (Figure S6). In
343 these cases, both samples contain similar proportions of false alleles from the 3rd ingroup population
344 and therefore no significant deviation from zero is observed. Nonetheless, if coverage between the
345 two tested samples is highly different (and thus one of the samples has a higher proportion of
346 misassigned reads), Z-scores become skewed and might be falsely interpreted as a signal of gene
347 flow (Figure S6). At the rate of index-hopping reported here (~0.470%), only genomes above 3X
348 coverage remain largely unbiased, and thus for ancient DNA studies where samples are being
349 multiplexed, elimination of index-hopping is of great importance.

350 We show that even with a low rate of index-hopping, such as the one observed in our empirical
351 study, downstream inferences can become biased if sample qualities are highly different. Therefore,
352 variation in sample endogenous DNA content are ideally kept to a minimum when sequencing a
353 sample pool on the same lane. Pre-pooling qPCR quantification of sample DNA (and endogenous
354 content) can be helpful to balance the sequencing libraries. Additionally, when multiplexed samples
355 are sequenced to high depth (i.e. across multiple lanes/flowcells), re-pooling could be considered
356 after the first sequencing run if high variation in (endogenous) read numbers is observed. This is

357 especially relevant for the NovaSeq platform, the most powerful sequencing platform currently
358 available, since it has been specifically designed for the multiplexing of up to hundreds of samples.

359 We show that in cases where low coverage data is generated or absolutely certainty is required, even
360 a low remaining rate of misassigned reads can cause severe downstream biases. For such studies we
361 therefore recommend the use of either short barcoded in-line adapters and/or dual indexing when
362 preparing pooled libraries for next generation sequencing, independently of sequencing platform.

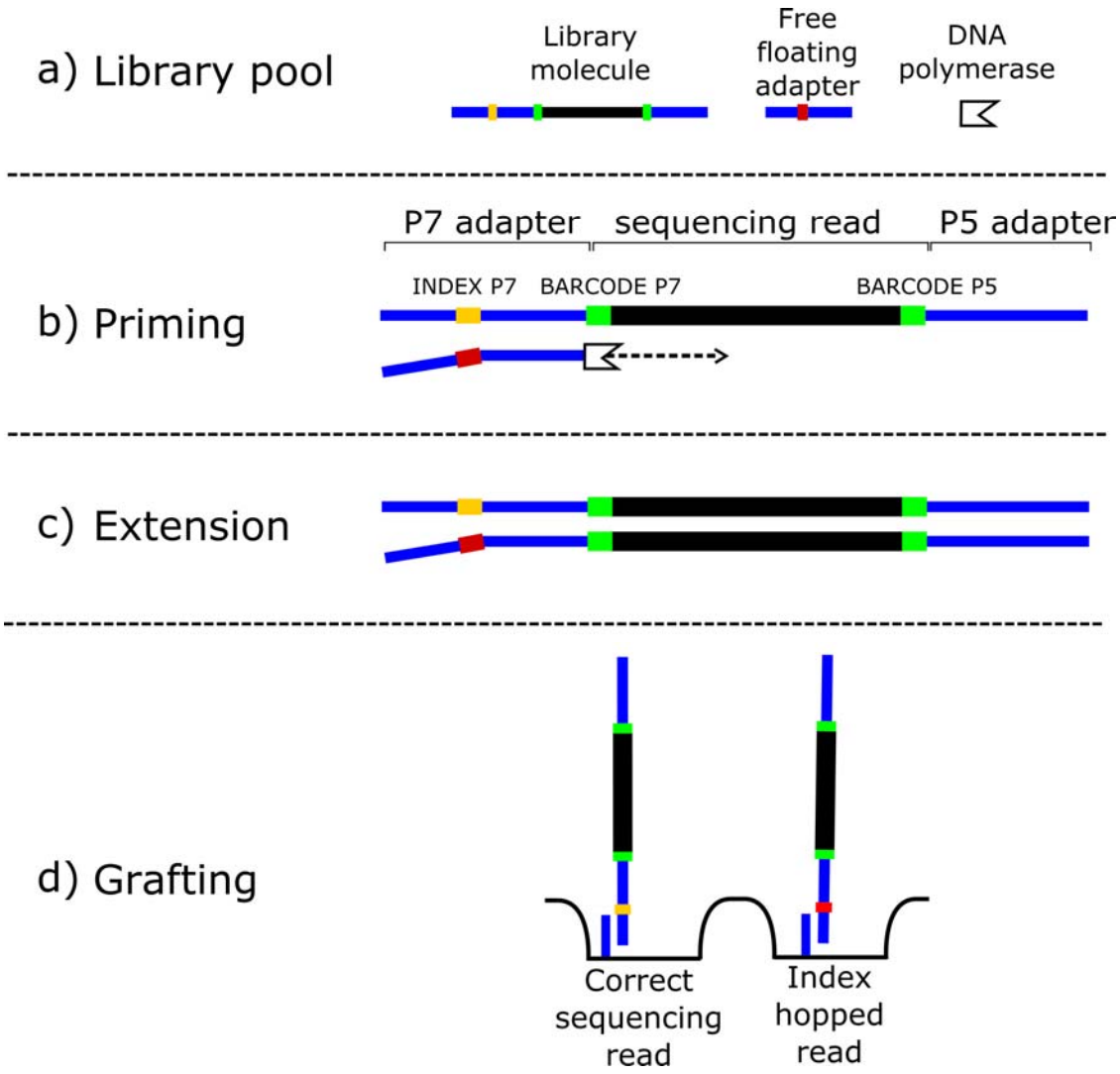

Figure 1: Schematic of index hopping during ExAmp clustering. A) The library pool, containing barcoded and indexed library molecules (black: DNA insert, green: P5 and P7 barcodes, orange: P7 index) and free-floating indexing primers, is mixed with ExAmp reagents before loading on the patterned flow cell. B) Free-floating adapter anneals to the adapter sequence of a library molecule and C) the library molecule gets extended by the DNA polymerase, forming a new library molecule with a wrong index. D) The library molecules are denatured, separating the strands, and each library molecule is allowed to graft onto a nanowell on the patterned flow cell.

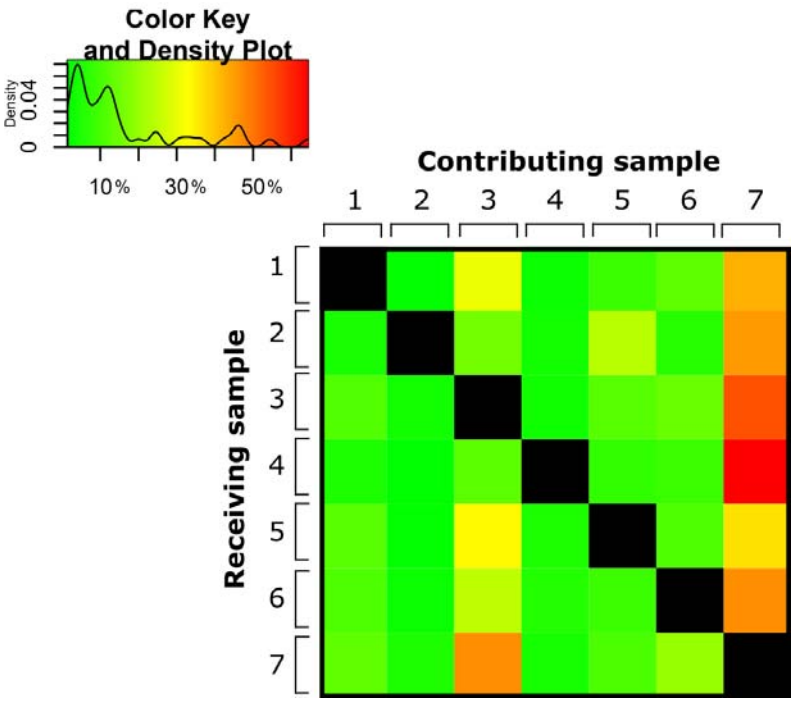

Figure 2: Proportion of hopped reads per sample out of all hopped reads. Samples in the top row contribute hopped reads, whereas samples on the left receive hopped reads. Samples with high number of sequenced reads (e.g. 3 and 7) are also the main contributors of hopped reads.

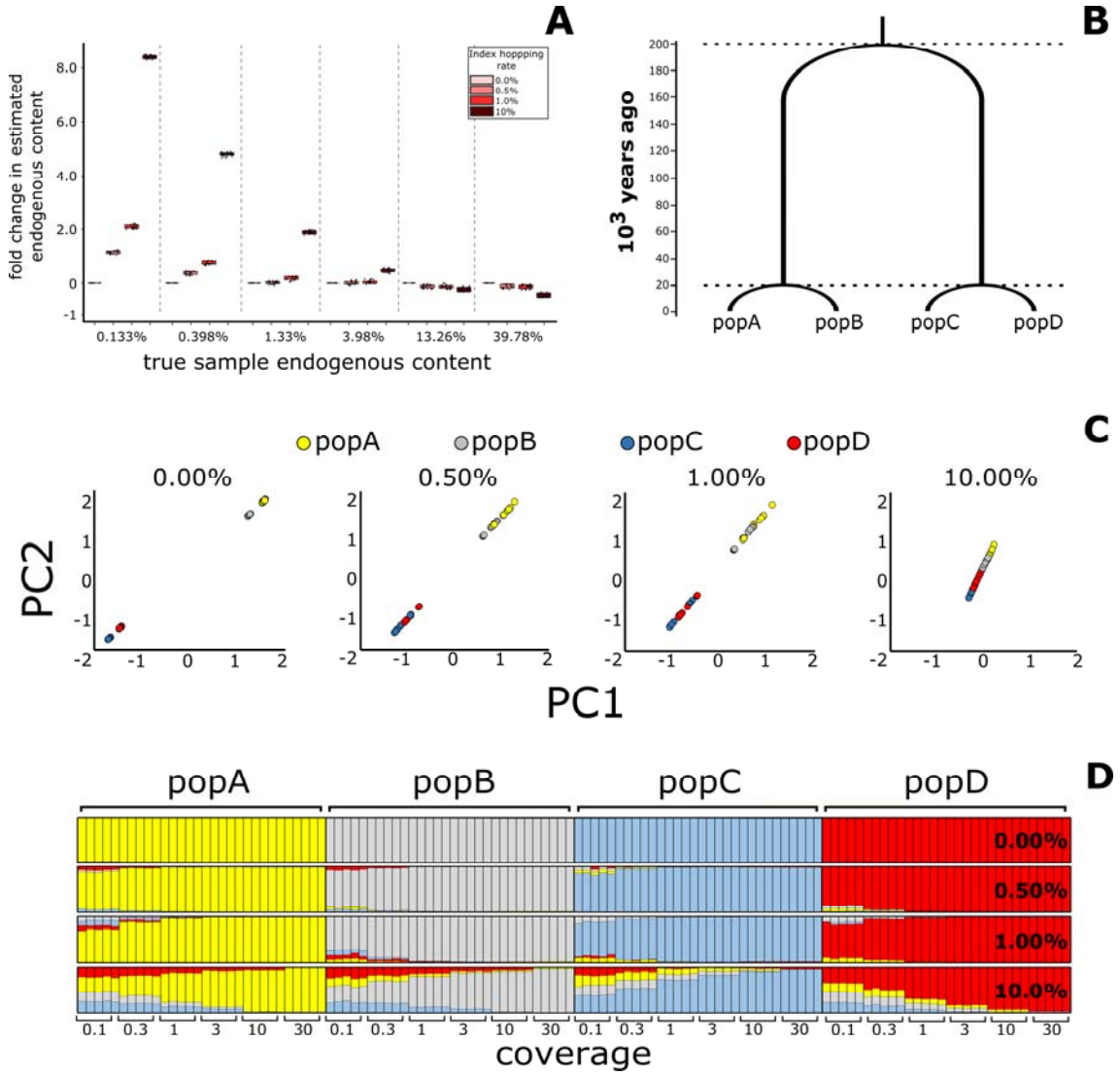

Figure 3: (A) Index hopping biases estimates of endogenous content. X-axis shows the simulated sample endogenous content, Y-axis shows the fold-change in the inferred endogenous content (fraction of mapped reads). Colours depict the simulated rate of index-hopping, which increases from left to right for each sample. Biases for samples with high endogenous content are minor (note that these samples appear to "loose" reads that are being assigned to samples of low endogenous content). However, samples of low quality (low endogenous content) are disproportionally affected by index hopping, leading to erroneously high estimates of endogenous DNA content. (B) Schematic representation of the simulated populations and their divergence times. For each populations 30 individuals with different levels of endogenous content are simulated. (C) Effects of index hopping on inferences of population differentiation: Principal Components Analysis for all individuals under different rates of index hopping (depicted on the top). Each plot shows all 120 individuals from the four simulated populations. (D) Effects of index hopping on inferences of individual admixture

388 *proportion. Each bar is an individual, X-axis depicts sample coverage. Percentage at the right depict*

389 *simulated index-hopping rate.*

390 *Table 1: Sequencing statistics and estimates of contamination and index hopping.*

Sequencing run	*Sequencing reads after quality filtering*	*Sequencing reads with wrong barcode pairs*	*Sequencing reads with wrong barcode pairs (%)*	*Sequencing reads with wrong index-barcode combination*	*Sequencing reads with wrong index-barcode combination (%)*
Run 1	316203540	99575	0.0301	1543241	0.488
Run 2	127766205	42457	0.0315	837926	0.656
Run 3	429511898	94527	0.0211	2740612	0.638
Average	291160548	78853	0.028	1707260	**0.594**

391

Author contributions: TvdV and MO performed the wetlab experiments. TvdV and FV performed computational data-analysis of the sequenced libraries. TvdV ran the simulations. TvdV and KG established the experimental design. TvdV, LD and KG conceived the study, interpreted the results and wrote the manuscript (with contribution from all authors).

Acknowledgments:

We acknowledge the support from the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. We thank Illumina for providing sequencing reagents. Illumina had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data access: Raw sequencing data is available at the European nucleotide archive under accession number XXXX.

All script that were used to simulate the data are available on github: https://github.com/XXXX

Funding sources: This project was supported by FORMAS grant 2015-676 to LD, FORMAS grant 2016-00835 to KG and the Jan Löfqvist Endowments of the Royal Physiographic Society of Lund to KG.

Conflicts of Interest

The authors declare no conflicts of interests.

411 **References**

412 Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH (2018) Direct estimation of
413 mutations in great apes reveals significant recent human slowdown in the yearly mutation rate.
414 *bioRxiv*, 287821.

415 Briggs AW, Stenzel U, Meyer M *et al.* (2010) Removal of deaminated cytosines and detection of in
416 vivo methylation in ancient DNA. *Nucleic acids research*, **38**, e87.

417 Carlsen T, Aas AB, Lindner D *et al.* (2012) Don't make a mista(g)ke: Is tag switching an overlooked
418 source of error in amplicon pyrosequencing studies? *Fungal Ecology*, **5**, 747–749.

419 Craig DW, Pearson J V, Szelinger S *et al.* (2008) Identification of genetic variants using bar-coded
420 multiplexed sequencing. *Nature Methods*, **5**, 887–893.

421 D'Amore R, Ijaz UZ, Schirmer M *et al.* (2016) A comprehensive benchmarking study of protocols and
422 sequencing platforms for 16S rRNA community profiling. *BMC genomics*, **17**, 55.

423 Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput
424 amplicon-sequencing. *Nucleic Acids Research*, **43**, 2513–2524.

425 Fumagalli M, Vieira FG, Korneliussen TS *et al.* (2013) Quantifying population genetic differentiation
426 from next-generation sequencing data. *Genetics*, **195**, 979–992.

427 Gordon D, Huddleston J, Chaisson MJP *et al.* (2016) Long-read sequence assembly of the gorilla
428 genome. *Science*, **352**, aae0344-aae0344.

429 Green RE, Krause J, Briggs AW *et al.* (2010) A Draft Sequence of the Neandertal Genome. *Science*,
430 **328**, 710–722.

431 Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC (2018) Detection and removal of barcode
432 swapping in single-cell RNA-seq data. *Nature Communications*, **9**, 2667.

433 Illumina Inc. NovaSeq 6000 Sequencing system.

434 Illumina Inc. (2017) *Effects of Index Misassignment on Multiplexing and Downstream Analysis*.

435 Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex
436 sequencing on the Illumina platform. *Nucleic acids research*, **40**, e3.

437 Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing
438 Data. *BMC Bioinformatics*, **15**, 356.

439 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
440 *Bioinformatics*, **25**, 1754–1760.

441 Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target
442 capture and sequencing. *Cold Spring Harbor Protocols*, **5**, pdb.prot5448.

443 Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids*
444 *Research*, **18**, 1687–1691.

445 Mitra A, Skrzypczak M, Ginalski K, Rowicka M (2015) Strategies for achieving high sequencing
446 accuracy for low diversity samples and avoiding sample bleeding using Illumina platform (C
447 Oudejans, Ed,). *PLoS ONE*, **10**, e0120520.

448 Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014) Analysis, Optimization and Verification
449 of Illumina-Generated 16S rRNA Gene Amplicon Surveys (MM Heimesaat, Ed,). *PLoS ONE*, **9**,
450 e94249.

451 Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by *Thermus*
452 *aquaticus* DNA polymerase I. *Nucleic acids research*, **23**, 2049–57.

453 Owens GL, Todesco M, Drummond EBM, Yeaman S, Rieseberg LH (2018) A novel post hoc method for
454 detecting index switching finds no evidence for increased switching on the Illumina HiSeq X.
455 *Molecular Ecology Resources*, **18**, 169–175.

456 Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J (2015) deML: robust demultiplexing of Illumina

457 sequences using a likelihood-based approach. *Bioinformatics (Oxford, England)*, **31**, 770–2.

458 Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D (2015) Partial uracil-DNA-glycosylase treatment
459 for screening of ancient DNA. *Philosophical transactions of the Royal Society of London. Series B,*
460 *Biological sciences*, **370**, 20130624.

461 Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed
462 target capture. *Genome Research*, **22**, 939–946.

463 Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets.
464 *Bioinformatics*, **27**, 863–864.

465 Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming,
466 identification, and read merging. *BMC Research Notes*, **9**, 88.

467 Shapiro B, Hofreiter M (2014) A paleogenomic perspective on evolution and gene function: New
468 insights from ancient DNA. *Science*, **343**, 1236573–1236573.

469 Sinha R, Stanley G, Gulati GS *et al.* (2017) Index Switching Causes "Spreading-Of-Signal" Among
470 Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*.

471 Skoglund P, Northoff BH, Shunkov M V *et al.* (2014) Separating endogenous ancient DNA from
472 modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of*
473 *Sciences of the United States of America*, **111**, 2229–34.

474 Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating Individual Admixture Proportions from
475 Next Generation Sequencing Data. *Genetics*, **195**, 693–702.

476 Smith AM, Heisler LE, St.Onge RP *et al.* (2010) Highly-multiplexed barcode sequencing: an efficient
477 method for parallel analysis of pooled samples. *Nucleic Acids Research*, **38**, e142–e142.

478 Team R Core (2016) R: A language and environment for statistical computing. R Foundation for
479 Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

480 van der Valk T, Sandoval-Castellanos E, Caillaud D *et al.* (2018) Significant loss of mitochondrial
481 diversity within the last century due to extinction of peripheral populations in eastern gorillas.
482 *Scientific Reports*, **8**, 6551.

483 Vodák D, Lorenz S, Nakken S *et al.* (2018) Sample-Index Misassignment Impacts Tumour Exome
484 Sequencing. *Scientific Reports*, **8**, 5307.

485 Wright ES, Vetsigian KH (2016) Quality filtering of Illumina index reads mitigates sample cross-talk.
486 *BMC genomics*, **17**, 876.