Estimating the rate of index hopping on the Illumina HiSeq X platform

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 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 potentially calls into question previously generated results and may make future use of the newest generation of platforms prohibitive. In this study we rely on barcodes, short sequences that are directly ligated to both ends of the DNA insert, which allows us to quantify the amount of index hopping. Correcting for multiple sources of noise, we identify on average only 0.470% of reads containing a hopped index. Multiplexing of samples on this platform is therefore unlikely to cause markedly different results to those obtained from older platforms.

Keywords

Read misassignment, next generation sequencing, ExAmp chemistry, multiplexing

Introduction

Multiplexing samples for next-generation sequencing is a common practice in many biological and medical applications [1–3]. During multiplexing, samples are individually labelled with unique identifiers (indices) that are embedded within one (single indexing) or both (dual indexing) sequencing platform-specific adapters [2,4]. The samples are subsequently pooled into a single DNA library and sequenced on the same lane, greatly reducing per sample sequencing cost. Following sequencing, computational demultiplexing, based on the sample-specific indices, allows for assignment of the sequenced reads to the respective sample of origin. However, ever since multiplexing approaches were introduced, low rates of read misassignment across samples sequenced on the same lane have been reported on all Illumina platforms [4–8]. Read misassignment is the results of reads carrying an unintended index and consequently being erroneously attributed to the wrong original sample. The reported rate of read misassignment on Illumina platforms that rely on the traditional bridge amplification for cluster generation is low (<1%) [9,10] and therefore this source of error has been mostly ignored.

However, on the latest generation of Illumina sequencing platforms (HiSeq X, HiSeq 4000 and NovaSeq) that rely on the exclusion amplification chemistry (ExAmp) in combination with patterned flow cells, a wide range of read misassignment rates has been reported [11–13]. Whereas one study reported no observable rate of read misassignment on neither HiSeq X and HiSeq 2500 platforms [14], other studies have documented rates of up to 10% [13]. According to Illumina's own estimates, the rate of read misassignment on platforms with ExAmp chemistry is up to 2% [9].

As a consequence of conflicting results, the prevalence and severity of read misassignment on the Illumina HiSeq X platforms remain unclear. This is partly due to the difficulties to reliably identify misassigned reads in sequencing experiments, particularly if pooling similar samples types (e.g. multiple individuals from the same population that have high sequence similarity). Recently, Illumina introduced dual indices on all ExAmp sequencing platforms, allowing for the filtering of the majority

of reads that show signs of read misassignment. However, since indices can potentially be switched at both ends of the molecule and the number of available indices is limited, it remains difficult to obtain direct estimates of read misassignment on these platforms.

Some research questions require high confidence in read identity, as presence of rare sequence variants can influence biological and medical conclusions. For instance, detection of low abundance transcripts or rare mutations can influence diagnostic inferences [11,15–17]. Studies with low input DNA quantities (e.g. single cell sequencing, ancient and historical DNA) are particularly susceptible to such errors [4] . Similarly, population genomics studies frequently rely on low-coverage genomic data, and presence of shared rare alleles across several populations or species can be interpreted as evidence for gene flow or admixture [18–21].

Processes resulting in read misassignment, i.e. presence of reads with a switched index, are numerous. The effect of sequencing errors that can convert one index sequence into another is well known and has led to series of recommendations for designing highly distinct indices [2]. Jumping PCR during bulk amplification of library molecules that carry different indices can generate chimeric sequences and should be avoided [22–25]. Similarly, cross-contamination of indexing adapters during oligonucleotide synthesis or laboratory work can lead to reads obtaining an unintended index. Additionally, cluster misidentification due to “bleeding” of indices into neighbouring clusters have been reported on all high throughput sequencing platforms [4–8,11,26]. However, for the latest Illumina platforms with patterned flow cells and ExAmp chemistry, read misassignment has been suggested to be caused by the presence of free-floating indexing primers in the final sequencing library [9,13]. Such free-floating molecules can appear if sequencing libraries are not stored properly and become fragmented or if final sequencing libraries are not properly size selected [27]. These primers and molecules can then anneal to the pooled library molecules and get extended by DNA polymerase before the rapid exclusion amplification on the flow cell, creating a new library molecule

with an erroneous index. We refer to this particular process of generating misassigned reads as index hopping.

In this study we make use of inline barcodes, short unique seven base pair sequences ligated to both ends of the DNA fragments [28], in combination with indexed primers that are traditionally used for sample identification. The barcodes become part of the sequencing read and thus allow for accurate identification of the read origin, even in the presence of index hopping. Therefore, the amount of index hopping can be directly quantified by identifying reads with wrong barcode-index combinations. We specifically use historical museum-preserved samples that are characterized by low DNA quantity and quality (the DNA is degraded, chemically modified and shows single-strand overhangs [29,30]). Libraries constructed from such low quality samples are prone to index hopping, since generally these samples yield low coverage sequencing data and inferences are based on subtle differences between limited sets of polymorphic sites [31,32]. Therefore, small quantities of misassigned DNA fragments in aDNA studies can already cause erroneous inferences [33] and it is thus crucial to distinguish genuine sample-derived ancient DNA fragments from false signals [34]. Additionally, estimates of ancient DNA preservation are often based on a small number of sequencing reads [35,36] and index hopping could therefore result in an erroneous inference about DNA preservation in a given sample. In the context of the study presented here, it is also worth noting that the small insert size in such aDNA samples allows us to obtain sequence information from both ends of the DNA fragment and thereby we can identify both barcodes with high accuracy.

Following sequencing on the HiSeq X platform, we identified a small fraction of reads (<1%) with a wrong barcode-index combination. After excluding several possible explanations, we conclude that index hopping happens in this system, but results in a similar rate of read misassignment as previously reported for older versions of Illumina sequencing platforms that rely on traditional bridge amplification for cluster generation. We therefore recommend using inline barcode-containing sequencing adapters, independent of the sequencing platform in studies that rely on low-coverage

102 data, require absolute certainty, aim to characterize rare variants or combine a large number of

103 samples that exceed available index combinations.

Methods

Library preparation and sequencing

DNA extracts from seven historical gorilla samples were turned into sequencing libraries following the strategy outlined in [28,37] (see supplementary material). All library preparation steps except indexing PCR were performed in a dedicated ancient DNA facility to minimize contamination. Briefly, 20 µl DNA extract was used in a 50 µl blunting reaction together with USER enzyme treatment to remove uracil bases resulting from aDNA damage [38]. DNA fragments within each sample were then ligated to a unique combination of incomplete, partially double-stranded P5- and P7-adapters, each containing a unique seven base pair sequence [37] (Table S1). We refer to these as the P5 and P7 barcodes from here on. All barcode sequences were at least three nucleotides apart from each other to ensure high certainty during demultiplexing and avoid converting one barcode into another through sequencing errors [37] (Table S1). To increase the complexity of the pooled sequencing library, one sample (sample 7) was split in two fractions, each of which received a different barcode combination (Table 1).

Indexing PCR was performed for 10 cycles using a unique P7 indexing primer for each sample, as in [2] (Table S1). We refer to the unique sequence added during the indexing PCR as the P7 index. As with the barcodes, all index sequences differed by at least three base pairs from each other (Table S1). Indexing PCR for sample 7 was performed in a single reaction combining both fractions of this sample. Following the indexing PCR, each DNA fragment contained three unique identifiers: the P5 and P7 barcodes directly ligated to the ends of the DNA fragments, and the P7 index, which becomes part of the Illumina sequencing adapter (Figure 1). Sample libraries were cleaned using MinElute spin columns, fragment length distribution and concentrations were measured on the Bioanalyzer. We then pooled all seven sample libraries in a ratio of 2:1:2:1:1:1:2 for samples 1 to 7, and performed two rounds of AMPure XP bead clean-up, using 0.5X and 1.8X bead:DNA ratio, respectively. We confirmed that indexing primers were successfully removed during clean-up by running the final

129 library on a Bioanalyzer (Figure S2). The pooled library was sequenced on three HiSeq X lanes that
130 were part of independent runs with a 5% phiX spike-in, at the SciLife sequencing facility in Stockholm.

131

132 **Data processing**

133 All reads were demultiplexed based on their unique indices using Illumina's bcl2fastq (v2.17.1)
134 software with defaults settings, allowing for one mismatch per index and only retaining "pass filter"
135 reads (Illumina Inc.). All unidentified reads, i.e. reads containing indices not used in our experiment,
136 were subjected to the same filtering steps, as described below. We removed adapter sequences
137 using AdapterRemoval V2.1.7 with standard parameters [39]. Due to the fragmented nature of DNA
138 in historical samples, we could subsequently merge the reads, requiring a minimal overlap of 11bp
139 and allowing for a 10% error rate. The merging of reads allows us to obtain sequencing information
140 for the complete DNA molecule and thus accurate identification of the barcodes on both ends of the
141 DNA fragment (P5 and P7 barcodes, respectively, Figure 1). Unmerged reads and reads below 29 bp
142 were removed. To increase certainty, we only retained reads with error-free P5 and P7 barcodes and
143 an average quality score of at least 30 using prinseq V0.20.4 [40].

144

145 **Disentangling cross-contamination from index hopping**

146 Low rates of cross-contamination between barcodes and indexes can be expected, even if strict
147 measure, such as clean-room facilities, are taken during library preparation [4]. This can result in
148 reads containing a wrong index-barcode pair and could be falsely taken as evidence for index
149 hopping. Since the inline barcodes used in this experiment are unaffected by index-hopping (Figure
150 1), we can accurately estimate the rate of barcode cross-contamination as the fraction of reads
151 containing a P5-P7 barcode pair that was not used during library preparation. In rare cases, barcode
152 cross-contamination results in a read with a valid barcode pair (e.g. a barcode combination that was

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

157 As the Illumina HiSeq X platform did not support a dual-indexing design at the time of this
158 experiment, estimating the rate of index cross-contamination could not be based on invalid index
159 pairs. Therefore, we relied on the fact that only seven out of the 40 indices that are routinely used in
160 our laboratory, were implemented in this experiment (Table S3). Assuming a relative equal rate of
161 cross-contamination between all 40 indexes, we estimated index cross-contamination as the fraction
162 of reads containing any of the 33 indices that were not included during library preparation.

163 We then determined the raw rate of index hopping as the fraction of reads showing an index-
164 barcode combination not used during the library preparation. We accounted for the possibility of
165 barcode and index cross-contamination resulting in the same barcode-index combination by
166 subtracting the contamination estimates obtained above from the raw value of index hopping. All
167 statistical analyses were performed in R 2.15.3 [41] (see supplementary material).

168

Results

Our sequencing libraries were made from degraded historical samples and thus contained a large proportion of short DNA fragments (Figure S3A). Therefore, the majority of reads could be confidently merged for all three sequencing runs (95.3% SE ± 1.0%). After all filtering steps (see Methods), the final dataset contained 89.3% SE ± 1.9% of the original sequence reads.

Barcode and index cross-contamination

We estimate the levels of barcode cross-contamination at 0.0276% SE ± 0.0026 across all three runs (Table 1, Table S2, Figure S1), with different rates observed between samples (global chi-square test, $P<10^{-15}$). The high observed level of incorrect barcode-pairings in sample seven, can be explained by formation of chimeric reads during pooled amplification of the two different fractions of this sample that were barcoded separately (see supplementary material). Assuming that adapter ligation of barcodes is unbiased with respect to the barcode sequence [37], this low percentage of cross-contamination will lead to a neglectable fraction of reads ($1.09 \cdot 10^{-8}$%, see supplementary material) with barcode pair that wrongly appear as having undergone index hopping.

The rate of index cross-contamination was estimated at 0.124% SE ± 0.0023 (Table S3), by quantifying the fraction of reads containing indices that were not intentionally used in our experiment (see Methods, Table S3).

Index hopping

Index hopping will not affect the barcodes that are directly ligated to the DNA fragments. Therefore, it can be readily distinguished from barcode cross-contamination by identifying reads containing a combination between an index and a barcode pair that was not used during the library preparation. Across all three sequencing runs, we detected a low proportion of such reads (mean=0.594%, SE ±

193 0.0434%, Table 1). As previously estimated ~0.124% of these reads are a result of index and barcode
194 cross-contamination. Therefore, the estimated rate of index hopping in our experiment across all
195 three sequencing runs is ~0.470% SE ± 0.044 (0.594% minus 0.124%). The proportion of hopped
196 reads differed significantly by sample (chi-square test, $P<10^{-15}$) and was positively correlated with the
197 number of sequenced reads per sample (Pearson's r = 0.96, P = 0.0005). This suggests that in
198 multiplexed sequencing runs the samples with higher number of sequenced reads will serve as the
199 dominant source of hopped reads (Figure 2). Even though the overall rate of index hopping is low,
200 samples with proportionally few sequenced reads are thus considerably more affected by index
201 hopping. In our experiments, this resulted in 2.49% SE ± 0.29% of index hopped reads in the sample
202 with the lowest number of sequenced reads (Table 1, Table S4, Figure 2).

203 In addition, the rate of index hopping differed significantly by read length and GC content. Reads
204 shorter than 90 bp and reads with GC content above 40% showed significantly higher proportion of
205 hopped reads than expected under a random distribution (chi-square test, both $P<10^{-15}$, Figure S3).

206

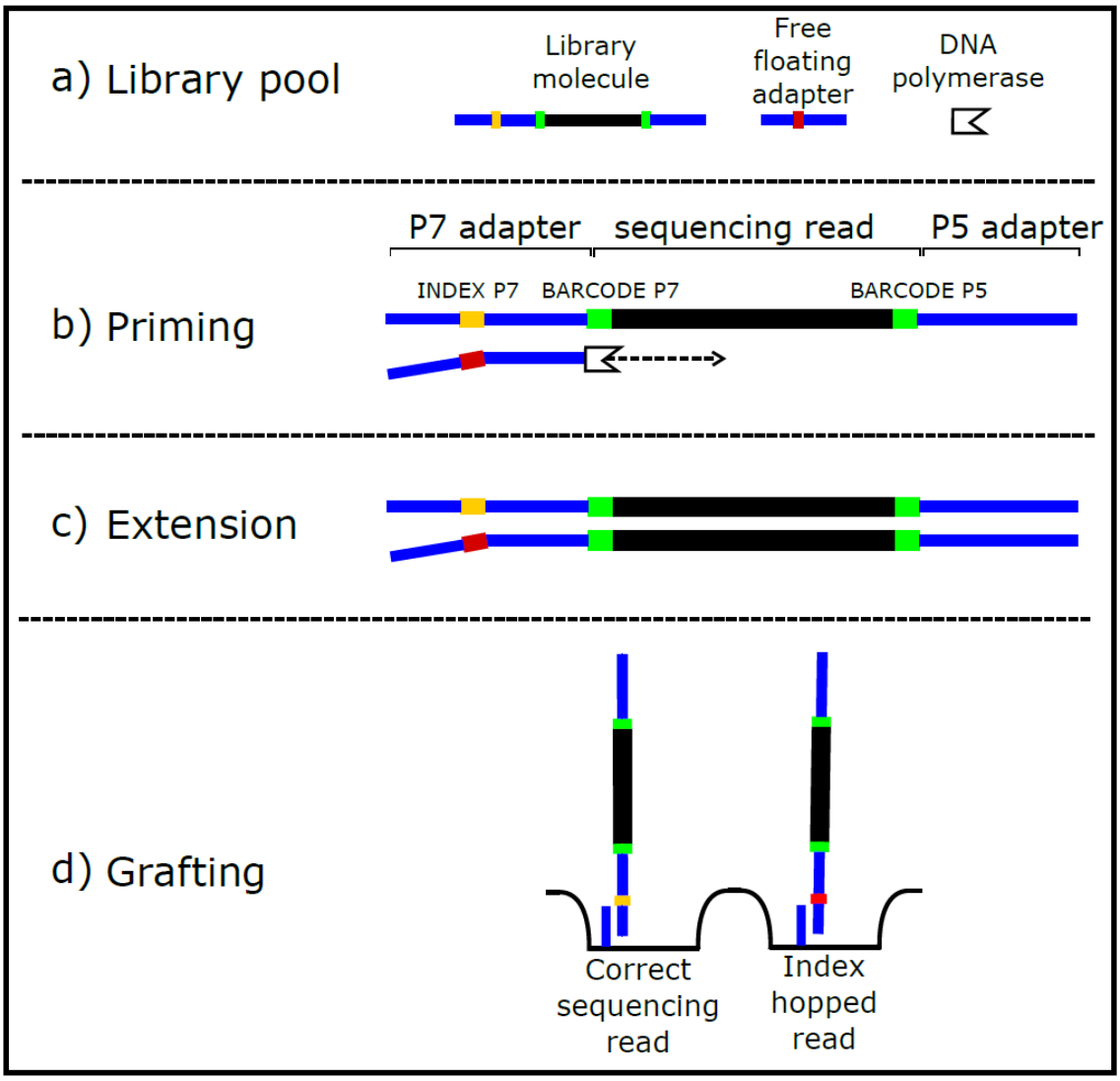


Figure 1: Outcome of index hopping. A) The library pool, containing barcoded and indexed library molecules and free-floating indexing primers, is mixed with ExAmp reagents before loading on the patterned flow cell. B) Free-floating adapters anneal to the adapter sequence of a library molecule and C) the library molecule subsequently gets extended by DNA polymerase forming a new library molecule containing 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.

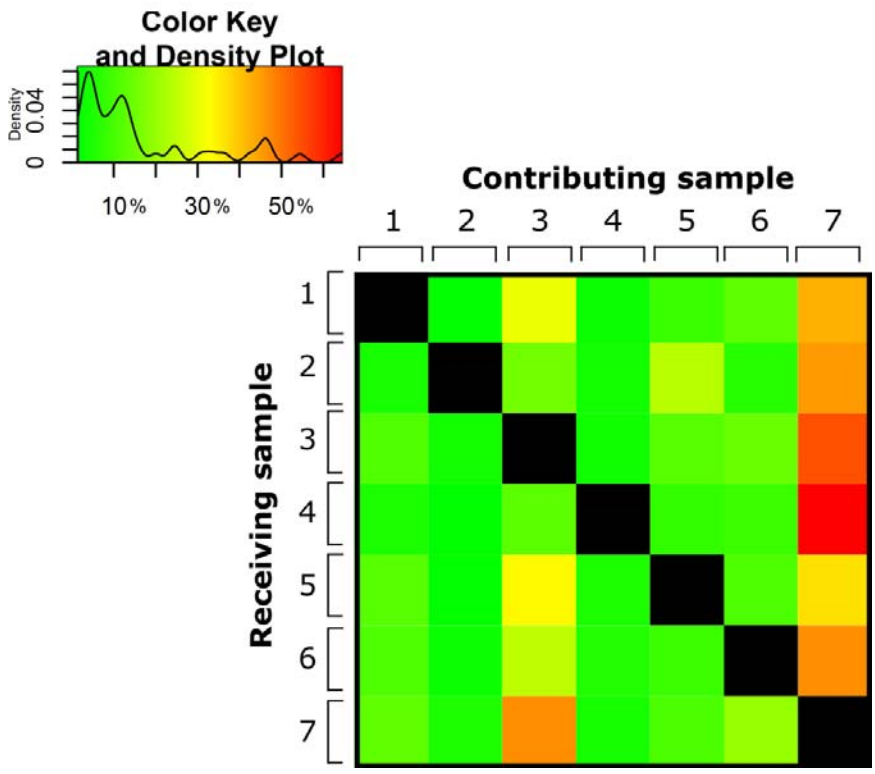


Figure 2, Proportion of hopped reads per sample out of all hopped reads. Samples in the top row contribute hopped reads, whereas samples on the left receive hopped reads. Samples with high number of reads (e.g. 3 and 7) are also the main contributors of hopped reads.

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

Sample	Used P5 barcode	Used P7 barcode	Original reads (Millions)			Reads after quality filtering (Millions)			Reads with wrong barcode pairs (cross-contamination)			Reads with wrong index-barcode combination			Cross contaminated reads (%)			Index hopped reads (%)		
			Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
1	3	3	40.63	14.74	53.95	34.49	13.01	50.43	2042	971	3580	158905	87697	280808	0.0059	0.0075	0.0071	0.4587	0.6693	0.5538
2	4	4	11.28	4.83	14.16	9.80	4.37	13.51	1447	644	1398	130125	77502	186512	0.0148	0.0147	0.0103	1.3100	1.7435	1.3622
3	5	5	127.21	44.98	157.84	104.79	40.40	147.48	12939	4184	6937	265347	163502	572787	0.0123	0.0104	0.0047	0.2526	0.4031	0.3869
4	6	6	13.80	4.53	18.76	11.31	3.88	17.05	1831	865	2314	262764	127417	349123	0.0162	0.0223	0.0136	2.2700	3.1824	2.0061
5	7	7	22.69	10.27	34.67	20.00	9.13	32.23	7555	3676	9563	166238	86408	308164	0.0378	0.0402	0.0297	0.8245	0.9374	0.9472
6	8	8	30.46	13.57	39.58	27.56	12.65	38.09	2034	1245	2449	78427	39006	161356	0.0074	0.0098	0.0064	0.2838	0.3073	0.4219
7	9 14	9 14	125.50	49.63	142.72	108.26	44.33	130.73	63867	26727	62677	481435	256394	881862	0.0590	0.0603	0.0479	0.4427	0.5751	0.6700
Unidentified	-	-	21.55	9.43	33.00	14.44	7.19	18.14	7860	4145	5609	-	-	-	0.0544	0.0576	0.0309	-	-	-
Total	-	-	393	152	495	331	135	448	99575	42457	94527	1543241	837926	2740612	-	-	-	-	-	-
Average barcode cross-contamination (%)	-	-	-	-	-	-	-	-	0.0301	0.0315	0.0211	-	-	-	-	-	-	-	-	-
Average index hopping (%)	-	-	-	-	-	-	-	-	-	-	-	0.488	0.656	0.638	-	-	-	-	-	-

Discussion

We show that index hopping occurs on the Illumina HiSeq X platform, but its rate is low in our study (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 we could systematically address each of them through a careful experimental design. Jumping PCR can be eliminated as explanation for wrong index-barcode combinations, as we avoided amplification of pooled libraries from different samples. 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 the observed proportion of reads with wrong index-barcode pairs.

Read misassignment is not a novel phenomenon for the Illumina sequencing platforms. Reported error rates range from 0.1% to 0.582% for HiSeq 2500 [4,6,42] and from 0.06% to 0.51% for the MiSeq platforms [5,7,8]. 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 HiSeq platforms with ExAmp chemistry [12–14]. 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 propose that the low rate of index hopping achieved in our experiment is the result of rigorous removal of free-floating adapters through size selection and library clean-up (Figure S2). Therefore, as previously suggested, strict library clean-up, such as those performed in this study, should be performed in multiplexed sequencing studies [27].

A novel observation in our study is that the number of hopped reads within a sample is proportional to the total number of reads contributed by the sample to the pooled sequencing library. Importantly, pooling samples in different quantities leads to a greater proportion of hopped reads into samples with fewer sequenced reads. In this study, libraries with the lowest number of sequenced reads (e.g. sample 2 and 4) displayed up to 3.2% of misassigned reads (Table 1), an order

of magnitude higher than the average rate within a lane. The effect of this skewed rate of index hopping becomes even more severe if the endogenous content is markedly different between samples, as is often observed in aDNA studies [43–45]. Since the endogenous content is usually not know beforehand, pooling samples in equal quantities can lead to large differences in the number of endogenous reads between samples. In these cases, the proportion of false assigned endogenous reads can reach rates above 10% (Figure S4). Therefore, variation in sample DNA quantities should be minimized when sequencing a sample pool on the same lane. Ideally, pre-pooling qPCR quantification of sample DNA should be performed to balance the sequencing libraries. Additionally, when samples are sequenced at high depth (i.e. across multiple lanes/flowcells), re-pooling should be considered after the first sequencing run if high variation in (endogenous) read numbers is observed. This is especially relevant for the NovaSeq platform, the most powerful sequencing platform currently available, since it has been specifically designed for the multiplexing of up to hundreds of samples. Additionally, we detected a higher rate of index hopping among the shorter reads and small differences in the fraction of index-hopped reads related to read GC-content. This suggests that the annealing of free floating adapters present in the sequencing libraries does not occur randomly. The underlying mechanisms are not yet well understood but could be related to differences in the DNA denaturation temperatures between DNA fragments of different size. Due to the lower denaturation temperature, short fragments might be occurring at a higher rate in single stranded conformation and are thereby more accessible to free floating index primers.

The recent addition of dual-indexing on the HiSeq X platform allows for the mitigation of the majority of errors resulting from index hopping. Nonetheless, in cases where low coverage data is generated or absolutely certainty is required, even a low remaining rate of misassigned reads might represent a major problem. Our study outlines a strategy for identification, quantification and removal of misassigned reads that result from index hopping or other sources of error. We therefore recommend the use of short barcoded in-line adapters when preparing pooled libraries for next

275 generation sequencing, independently of sequencing platform, if a very high degree of certainty is

276 required.

277 **Author contributions:** TvdV and MO performed the wetlab experiments. TvdV and FV performed
278 computational data-analysis. TvdV and KG established the experimental design. TvdV, LD and KG
279 conceived the study, interpreted the results and wrote the manuscript (with contribution of all co-
280 authors).

281

282 **Acknowledgments:**

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

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

291 **Funding sources:** This project was supported by FORMAS grant 2015-676 to LD, FORMAS grant 2016-
292 00835 to KG and the Jan Löfqvist Endowments of the Royal Physiographic Society of Lund to KG.

293 **Conflicts of Interest**

294 The authors declare no conflicts of interests.

295

296 **References**

297 1. Craig, D.W., Pearson, J. V, Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski,
298 T.L., Laub, T., Nunn, G., Stephan, D.A., *et al.* (2008). Identification of genetic variants using
299 bar-coded multiplexed sequencing. Nat. Methods *5*, 887–893.

300 2. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly
301 multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *5*, pdb.prot5448.

302 3. Smith, A.M., Heisler, L.E., St.Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris,
303 A.N., Perry, K.M., Giaever, G., Pourmand, N., *et al.* (2010). Highly-multiplexed barcode
304 sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. *38*,
305 e142–e142.

306 4. Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in
307 multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

308 5. Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., and Kelso, J. (2015). deML: robust
309 demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics *31*,
310 770–2.

311 6. Wright, E.S., and Vetsigian, K.H. (2016). Quality filtering of Illumina index reads mitigates
312 sample cross-talk. BMC Genomics *17*, 876.

313 7. Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., and Graf, J. (2014). Analysis,
314 Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. PLoS
315 One *9*, e94249.

316 8. D'Amore, R., Ijaz, U.Z., Schirmer, M., Kenny, J.G., Gregory, R., Darby, A.C., Shakya, M., Podar,
317 M., Quince, C., and Hall, N. (2016). A comprehensive benchmarking study of protocols and
318 sequencing platforms for 16S rRNA community profiling. BMC Genomics *17*, 55.

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

10. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. - Supplement. Nature *456*, 53–9.

11. Vodak, D., Lorenz, S., Nakken, S., Aasheim, L.B., Holte, H., Bai, B., Myklebost, O., Meza-Zepeda, L.A., and Hovig, E. (2017). Sample-Index Misassignment Impacts Tumor Exome Sequencing. bioRxiv.

12. Griffiths, J.A., Lun, A.T.L., Richard, A.C., Bach, K., and Marioni, J.C. (2017). Detection and removal of barcode swapping in single-cell RNA-seq data. bioRxiv.

13. Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T., Morganti, R.M., *et al.* (2017). Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. bioRxiv.

14. Owens, G.L., Todesco, M., Drummond, E.B.M., Yeaman, S., and Rieseberg, L.H. (2017). A Novel Post Hoc Method For Detecting Index Switching Finds No Evidence For Increased Switching On The Illumina HiSeq X. bioRxiv.

15. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. U. S. A. *109*, 14508–13.

16. Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H.P. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res. *40*, e2–e2.

17. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol.

31, 46–53.

18. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W.W., Fritz, M.H.Y., *et al.* (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.

19. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One *7*, e37558.

20. Fumagalli, M., Vieira, F.G., Korneliussen, T.S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. Genetics *195*, 979–992.

21. Therkildsen, N.O., and Palumbi, S.R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol. Ecol. Resour. *17*, 194–208.

22. Meyerhans, A., Vartanian, J.P., and Wain-Hobson, S. (1990). DNA recombination during PCR. Nucleic Acids Res. *18*, 1687–1691.

23. Odelberg, S.J., Weiss, R.B., Hata, A., and White, R. (1995). Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. Nucleic Acids Res. *23*, 2049–57.

24. Carlsen, T., Aas, A.B., Lindner, D., Vrålstad, T., Schumacher, T., and Kauserud, H. (2012). Don't make a mista(g)ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecol. *5*, 747–749.

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

26. Mitra, A., Skrzypczak, M., Ginalski, K., and Rowicka, M. (2015). Strategies for achieving high

366 sequencing accuracy for low diversity samples and avoiding sample bleeding using Illumina
367 platform. PLoS One *10*, e0120520.

368 27. Illumina Inc. (2017). Effects of Index Misassignment on Multiplexing and Downstream
369 Analysis.

370 28. Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries
371 for multiplexed target capture. Genome Res. *22*, 939–946.

372 29. Mulligan, C.J. (2005). Isolation and Analysis of DNA from Archaeological, Clinical, and Natural
373 History Specimens. Methods Enzymol. *395*, 87–103.

374 30. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Paabo, S. (2012). Temporal patterns
375 of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One *7*, e34131.

376 31. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J.,
377 Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., *et al.* (2015). Population genomics of
378 Bronze Age Eurasia. Nature *522*, 167–172.

379 32. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G.,
380 Nordenfelt, S., Harney, E., Stewardson, K., *et al.* (2015). Massive migration from the steppe
381 was a source for Indo-European languages in Europe. Nature *522*, 207–211.

382 33. Wall, J.D., and Kim, S.K. (2007). Inconsistencies in Neanderthal Genomic DNA Sequences. PLoS
383 Genet. *3*, e175.

384 34. Skoglund, P., Northoff, B.H., Shunkov, M. V, Derevianko, A.P., Pääbo, S., Krause, J., Jakobsson,
385 M., and Klein, R.G. Separating endogenous ancient DNA from modern day contamination in a
386 Siberian Neandertal.

387 35. Slon, V., Hopfe, C., Weiß, C.L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., Rosas, A.,
388 Soressi, M., Knul, M. V, Miller, R., *et al.* (2017). Neandertal and Denisovan DNA from

Pleistocene sediments. Science *356*, 605–608.

36. Pečnerová, P., Díez-del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., Nikolskiy, P., Tikhonov, A., Vartanyan, S., and Dalén, L. (2017). Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth. Curr. Biol. *27*, 3505–3510.e3.

37. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *370*, 20130624.

38. Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. *38*, e87.

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

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

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

42. Wright, E.S., and Vetsigian, K.H. (2016). Inhibitory interactions promote frequent bistability among competing bacteria. Nat. Commun. *7*, 11274.

43. Damgaard, P.B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., and Allentoft, M.E. (2015). Improving access to endogenous DNA in ancient bones and teeth. Sci. Rep. *5*, 11184.

44. Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen,

F., Moiseyev, V., Gromov, A., Raczky, P., *et al.* (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. PLoS One *10*, e0129102.

45. van der Valk, T., Lona Durazo, F., Dalén, L., and Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. Mol. Ecol. Resour. *17*, e111–e121.

Supplementary materials for:

Estimating the rate of index hopping on the Illumina HiSeq X platform

Tom van der Valk, Francesco Vezzi, Mattias Ormestad, Love Dalén, Katerina Guschanski

This PDF includes:

Methods

Library preparation and sequencing

DNA extracts from seven historical eastern gorilla samples that previously yielded good sequencing results on the Illumina HiSeq 2500 platform and showed high endogenous content were turned into sequencing libraries following the strategy outlined in (1,2) as detailed below. All library preparation steps except indexing PCR were performed in a dedicated ancient DNA facility to minimize contamination.

1 USER-TREATMENT AND BLUNT-ENDING

We used 20 µl DNA extract in a 50 µl blunting reaction together with USER enzyme treatment to remove uracil bases resulting from aDNA damage (3) as follows:

	Volume per sample	*Final concentration*
H_2O	9.89 µl	
Tango Buffer (10x)	4 µl	1x
dNTPs (25 mM each)	0.16 µl	100 µM each
ATP (100 mM)	0.4 µl	1 mM
T4 PNK (10 U/µL)	2 µl	0.5 U/ µl
USER enzyme (1U/ µL)	2.75 µl	0.06U/ µl
T4 DNA Pol (5 U/µL)	0.8 µl	0.1 U/ µl
DNA	20 µl	
TOTAL VOLUME	40 µl	

Samples were incubated for 3 h at 37°C, followed by the addition of 1 µl T4 DNA polymerase (Thermo Scientific) and incubation at 25°C for 15 min and 12°C for 5 min. We then MinElute purified the reaction according to the manufacturer's protocol and eluted in 22 µl of EB buffer. DNA fragments within each sample were then ligated to a unique combination of incomplete, partially double-stranded P5- and P7-adapters (10 µM each), each containing a unique seven base pair sequence as follows:

450 **2) ADAPTER LIGATION AND BARCODING**

451

	Volume per sample	*Final concentration*
H_2O	10 µl	
T4 DNA ligase Buffer (10x)	4 µl	1x
PEG-4000 (50%)	4 µl	5%
Barcoded adapters (10 µM)	1 µl each	1.25 µM
T4 DNA ligase (5 U/ul)	1 µl	0.125 U/µl
Blunt-ended DNA	20 µl	
TOTAL VOLUME	40 µl	

452 **1 µl of the barcode P5.F-mix (10 µM) and 1 µL of the barcode P7.F-mix (10 µM).*

453 Samples were incubated for 30 minutes at room temperature. We refer to the 7 basepair adapters as
454 the P5 and P7 barcodes from here on. All barcode sequences were at least three nucleotides apart
455 from each other to ensure high certainty during demultiplexing and avoid converting one barcode
456 into another through sequencing errors (2, Table S1). To increase the complexity of the pooled
457 sequencing library, one sample received two different barcode combinations (Table 1). Following
458 adapter ligation reactions were MinElute purified according to the manufacturer's protocol and
459 eluted in 22 µl of EB buffer. We then performed adapter fill-in as follows:

460

461 **3) ADAPTER FILL-IN**

	Volume per sample	*Final concentration*
H_2O	14.1 µl	
Isothermal Amp. buffer 10x	4 µl	1x
dNTPs (25 mM each)	0.4 µl	250 uM each
Bst polymerase 2.0, Large Fragment (8 U/µL)	1.5 µl	0.4 U/µL
Adapter-ligated DNA	20 µl	
TOTAL VOLUME	40 µl	

462

463

After setting up the reaction samples were incubated for 20 min. at 37°C, followed by 20 min. at 80°C to inactivate the Bst polymerase and cleaned using MinElute spin columns as above. Indexing PCR was performed for 10 cycles in 125 µl volume using a unique P7 indexing primer for each sample, as follows:

PCR cocktail	*Volume **per 15 µL** DNA*	*Final concentrations*
H_2O	84 µL	
Pfu Turbo Cx Hotstart DNA Buffer 10x	12.5 µL	1x
dNTPs (25 mM each)	1.25 µL	250 µM each
Indexing primer P7 (10 µM)	5 µL	0.4 µM
Primer IS4 (10 µM)	5 µL	0.4 µM
Pfu Turbo Cx Hotstart DNA Polymerase (2.5 U/µL)	2.5 µL	5 U
Adapter-ligated DNA	15 µL	
TOTAL VOLUME	125 µL	

Cycling conditions:

2 min. 95°C, (95°C- 30sec. 59°C- 30sec. 72°C- 1min), 7min. 72°C, hold 8°C

474 **Data processing**

475 All reads were demultiplexed based on their unique indices using Illumina's bcl2fastq (v2.17.1)
476 software with defaults settings, allowing for one mismatch per index and only retaining "pass filter"
477 reads (Illumina Inc.). All unidentified reads, i.e. reads with indices that were not used in our
478 experiment, were subjected to the same filtering steps, as described below. We removed adapter
479 sequences using AdapterRemoval V2.1.7 using standard parameters and subsequently merged the
480 reads, requiring a minimal overlap of 11bp and allowing for a 10% sequencing error rate ([4]).
481 Unmerged reads and reads below 29 bp were removed leaving only merged reads with an original
482 insert size of at least 15 bp (7 bp barcodeP7 + 7 bp barcodeP5 + 15 bp DNA fragment = 29 bp). To
483 increase certainty, we only retained reads with intact and error-free P5 and P7 barcodes (assessed
484 using an in-house python script available upon request) and an average quality score of at least 30
485 using prinseq V0.20.4 ([5]).

486

487 **Statistical analyses**

488 Statistical analyses were performed in R 2.15.3 ([6]). Significant global chi-square tests were followed
489 by a post hoc procedure as implemented in the R package polytomous
490 (https://artax.karlin.mff.cuni.cz/r-help/ library/polytomous/html/00Index.html). The minimum value
491 of the chi-squared test statistic for the given degrees of freedom was used to assess if individual
492 observed values differ significantly from an overall hypothetical homogeneous distribution. The test
493 also identified the direction of these differences.

494

Estimating barcode cross-contamination

We estimated the rate of barcode cross-contamination as the fraction of reads containing a P5-P7 barcode pair that was not used during library preparation. There is a small probability that cross-contaminating changes both the P5 and P7 barcode into another valid barcode pair (e.g. a pair that was used during library preparation). The chance of this occurring is proportional to the number of used barcodes and can be estimates as $((n-1)\cdot\left(\frac{x}{n-1}\right)^2$, where *n* represents the number of used barcodes (eight in this study) and *x* the estimated average barcode cross contamination). The fraction of reads with invalid barcode-pairs was 0.0276%, which results in an estimate of $(8-1)\cdot\left(\frac{0.000276}{8-1}\right)^2\cdot 100\% = \mathbf{1.09\cdot 10^{-8}}\,\%)$ of undetected cross-contaminated reads.

Difference in barcode cross-contamination between samples

The rate of barcode cross-contamination differed significantly by sample (global chi-square test, $P<10^{-15}$). The implemented posthoc procedure suggested that samples 5 and 7 had significantly more reads with wrong barcode combinations than expected, whereas all the other samples had significantly fewer such reads. Among reads with barcode cross contamination we found an overrepresentation of incorrectly paired barcodes #9 and #14 (Table S2), both of which were used for sample 7 in the following combinations: P5-#9 with P7-#9 and P5-#14 with P7-#14 (Table 1). Elevated cross-contamination between these two barcodes during laboratory procedures could explain the results. However, the observed high rate of wrong barcode pairs (P5-#9 with P7-#14, P5-#14 with P7-#9, Figure S3) is more likely the result of jumping PCR during the 10 rounds of indexing PCR, as both fractions of sample 7 were indexed in a pooled reaction. Equal frequency of wrong barcode pairs is further supporting this notion (Table S2) and can be explained by jumping PCR happening randomly among the reads. In contrast, it is rather unlikely that all four barcodes would have received equal amounts of cross-contamination during laboratory procedures.

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

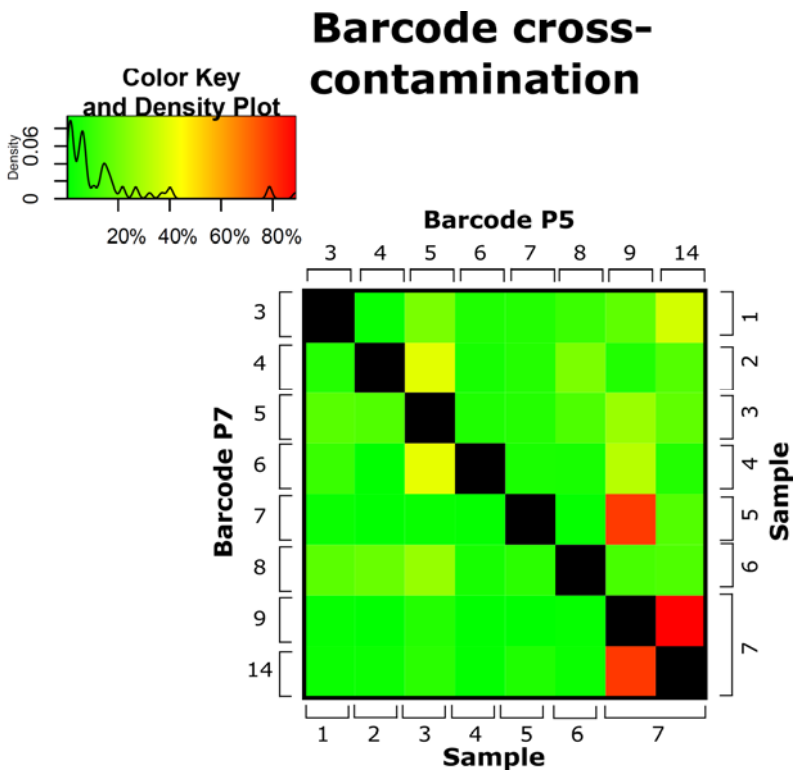
2. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** 20130624 (2015).

3. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38,** e87 (2010).

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

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

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



537

538 *Figure S1: Proportion of a given wrong barcode pair in the data out of all cross-contaminated barcode*

539 *pairs. Note that the overall rate of barcode cross-contamination is only* 0.0276% (Table 1) *Barcodes 9*

540 *and 14 are paired significantly more often due to the formation of chimeric reads during pooled*

541 *amplification of two fractions of sample 7 (see methods).*

542

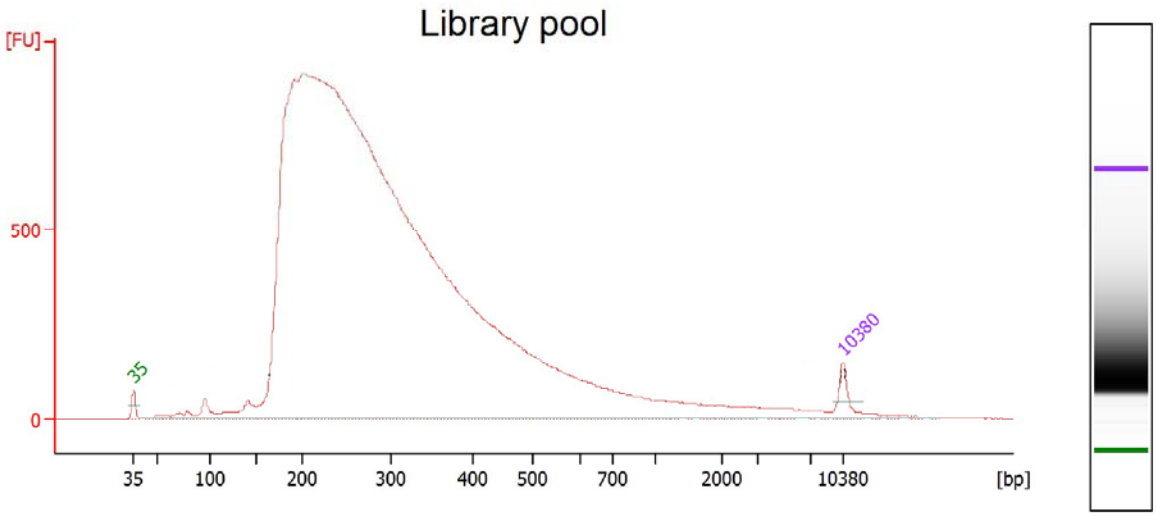
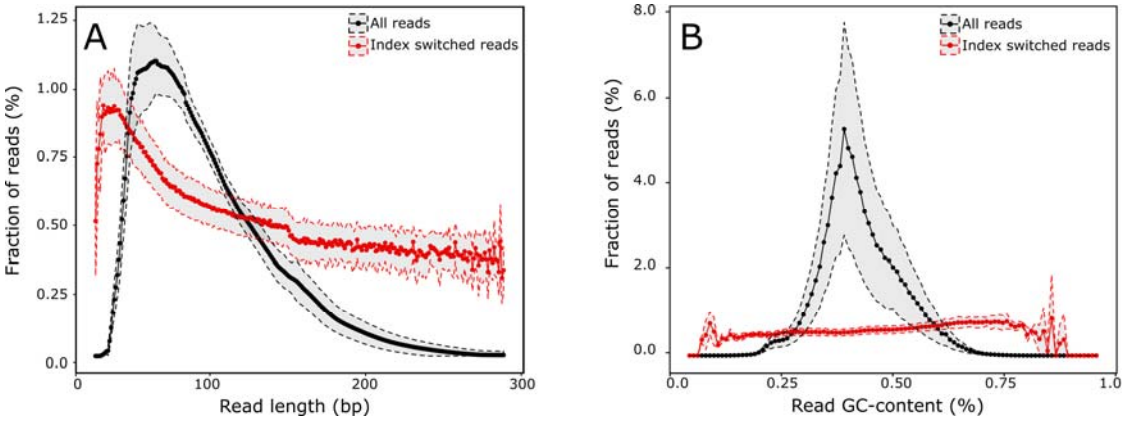


Figure S2: Bioanalyzer profile of the final pooled library. Note that during library preparation, sequencing adapters are attached to the DNA fragments, adding an additional 136 bp to the original DNA fragments. The insert size of the DNA is therefore 136 bp lower than what the Bioanalyzer shows.



549

550 *Figure S3: A) Read length distribution and the proportion of index hopping by read length. B) Read*

551 *GC-content distribution and the proportion of index hopping by read GC content. Shaded area depicts*

552 *95% confidence interval.*

553

554

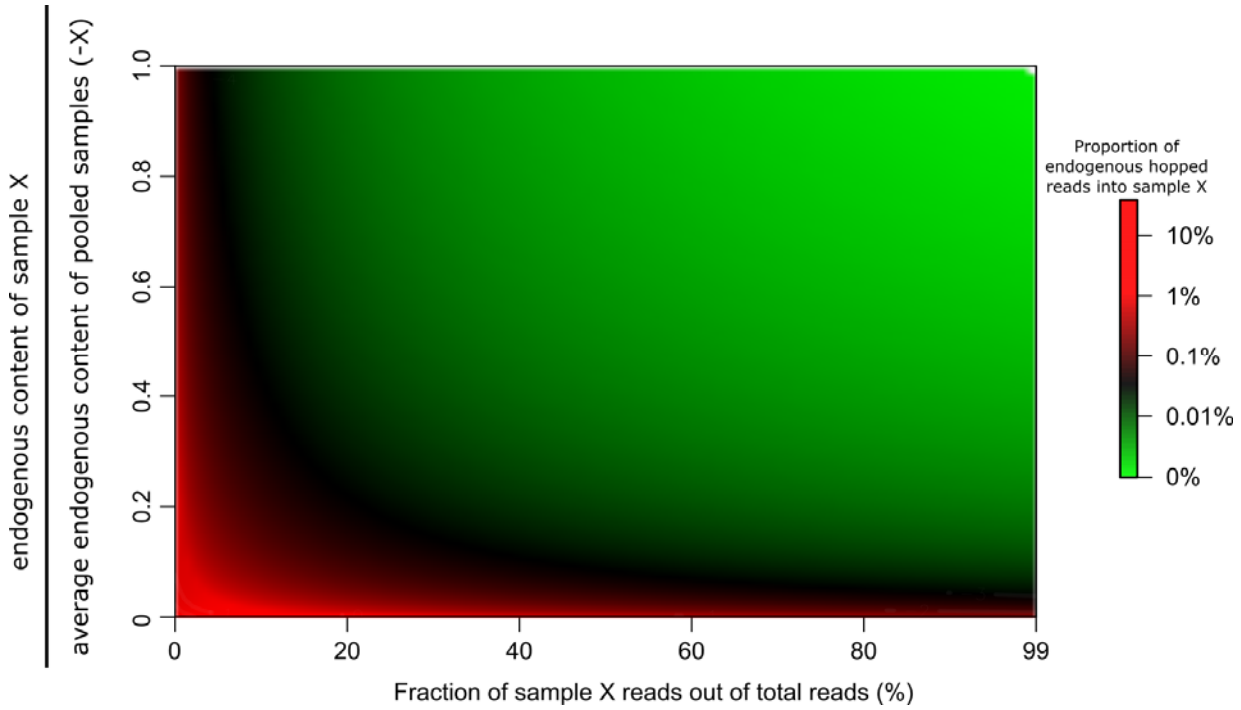


Figure S4. Theoretical relationship between endogenous content, fraction of total reads contributed by a given sample (referred to as sample X) to the pooled sequencing library, and index hopping. The lower the proportion of reads coming from sample X and the lower its endogenous content compared to other samples in the pooled sequencing library, the higher the proportion of endogenous hopped reads that sample X will receive from other samples.