Low 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

Keywords: Read misassignment, index hopping, inline barcodes, historical samples, ancient DNA

Abstract

The high throughput capacities of the Illumina sequencing platforms and the possibility to label samples with unique identifiers has encouraged a wide use of sample multiplexing. However, this practice results in low rates of read misassignment (<1%) across samples sequenced on the same lane on all Illumina sequencing platforms that rely on the traditional bridge amplification. Alarmingly high rates of read misassignment of up to 10% were recently reported for the newest Illumina machines (HiSeq X and HiSeq 4000). This potentially calls into question previously generated and published results and may make future use of these platforms prohibitive for many applications in biology and medicine. In this study we rely on inline barcodes that are ligated to both ends of the DNA insert, to directly quantify the amount of index hopping in historical museum-preserved samples. As the barcodes become part of the sequencing read, they allow us to reliably infer the read origin even in the presence of index hopping. After sequencing the same pooled library of seven samples on three independent HiSeq X lanes and accounting for multiple possible sources of error, including barcode and index cross-contamination, we identified on average only 0.470% hopped reads. We conclude that index hopping happens on the newest generation of Illumina sequencing platforms, but results in a similar rate of read missagnment as reported for older Illumina machines. We nonetheless recommend using inline barcodes in multiplexing studies that rely on low-coverage data, require absolute certainty and/or aim to characterize rare variants.

Introduction

Multiplexing samples for next-generation sequencing is a common practice in many biological and medical applications (Craig et al. 2008; Meyer and Kircher 2010; Smith et al. 2010; Caporaso et al. 2012; Rohland and Reich 2012). The high throughput capacities of most sequencing platforms clearly encourage multiplexing and optimized sequencing protocols with greater data output are continuously being developed. During multiplexing, samples are individually labelled with unique identifiers (indices) that are frequently embedded within one or both sequencing platform-specific adapters and are separated from the actual template (Meyer and Kircher 2010; Kircher et al. 2012, TruSeq Nano DNA kit (Illumina), NEBNext Ultra DNA kit (New England Biolabs)). The samples are subsequently pooled into a single sequencing library and sequenced on the same lane. 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 were reported on all Illumina platforms (Kircher et al. 2012; Nelson et al. 2014; D'Amore et al. 2016; Wright and Vetsigian 2016b), the most frequently used next generation sequencing technology (Research & Markets 2017). This process results in reads from one sample carrying a wrong index and consequentially being erroneously attributed to the wrong original sample. The reported rate of read misassignment is low (<1%) on Illumina platforms that rely on the traditional bridge amplification for cluster generation (Illumina Inc. 2017) and therefore this source of error has been readily ignored.

The use of the exclusion amplification chemistry (ExAmp) in combination with patterned flow cells on the newest generation of the Illumina sequencing platforms (HiSeq X and HiSeq 4000) was an important improvement, as it significantly increased data throughput and lowered sequencing cost (Illumina Inc. 2017). However, recently reported high rates of read misassignment of up to 10% observed for single cell RNA libraries sequenced on the Illumina HiSeq 4000 platform (Griffiths et al.

2017; Sinha et al. 2017) have shaken the scientific community, potentially calling into question many generated and published results. This finding is particularly worrying in light of the recently introduced NovaSeq sequencing platform, which offers even higher throughput while relying on the same technology as HiSeq X and HiSeq 4000. As even more samples can be multiplexed on a single lane, the potential bias from read misassignment would be further increased.

Several different processes can lead to read misassigment, i.e. presence of reads with a switched index. 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 (e.g. Meyer and Kircher 2010). Jumping PCR during bulk amplification of library molecules that carry different indices can generate chimeric sequences and should be avoided (Meyerhans et al. 1990; Odelberg et al. 1995; Lahr and Katz 2009; Holcomb et al. 2014; McDevitt et al. 2016). Similarly, cross-contamination of indexing adapters during oligonucleotide synthesis or laboratory work can lead to reads being attributed to the wrong sample of origin. Mixed clusters that can form on the flow cell if colonies from different template molecules grow into each other during cluster generation were identified as source of misassigned reads on older Illumina platforms (Kircher et al. 2012). For the Illumina platforms with patterned flow cells and ExAmp chemistry, read misassigment has been suggested to be caused by the presence of free-floating indexing primers in the final sequencing library. These primers can 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 a wrong index (Illumina Inc. 2017; Sinha et al. 2017). We refer to this particular process of generating misassigned reads as index hopping.

The preprint by Sinha and colleagues (2017) has started an active discussion about the prevalence of index hopping on the Illumina platforms with ExAmp chemistry. Illumina acknowledged a higher rate of index hopping on platforms with ExAmp chemistry compared to platforms that rely on bridge amplification for cluster generation, reporting up to 2% compared to ≤1% read misassignment (Illumina Inc. 2017). However, another study found no evidence for index hopping neither on HiSeq X nor on HiSeq 2500 platforms (Owens et al. 2017). Rigorously removing free-floating primers and adapters during library preparation by means of size-specific library clean-up was suggested to be the most efficient way to avoid index hopping (Illumina Inc. 2017; Griffiths et al. 2017; Sinha et al. 2017).

Due to the conflicting reports, the prevalence and severity of index hopping on Illumina HiSeq X and HiSeq 4000 platforms remain unclear. This is partly due to the difficulties to reliably identify missassigned reads in sequencing experiments, particularly if pooling similar samples types (e.g. multiple individuals from the same population that have high sequence similarity). However, some research questions clearly 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 (Greenman et al. 2007; Schmitt et al. 2012; Flaherty et al. 2012; Trapnell et al. 2013). Studies with low input DNA quantities are particularly susceptible to such errors. Besides single cell RNA sequencing, these include ancient and historical samples (Kircher et al. 2012). 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 (Green et al. 2010; Nielsen et al. 2012; Fumagalli et al. 2013; Allentoft et al. 2015; Wall et al. 2016; Therkildsen and Palumbi 2017).

In this study we make use of inline barcodes, short unique 7-bp sequences ligated to both ends of the DNA fragments (Rohland and Reich 2012), in combination with indexed primers that subsequently were used to amplify the libraries. This enabled us to directly quantify the amount of index hopping in historical museum-preserved samples. These barcodes become part of the sequencing read and thus allow for identification of the read origin, even in the presence of index hopping. Historical samples are characterized by low DNA quantity and quality (the DNA is degraded, chemically modified and shows single-strand overhangs (Mulligan 2005; Sawyer et al. 2012)). We purposefully use this low-quality sample source, as it has been suggested that libraries constructed from difficult samples may be more prone to index hopping than libraries constructed from high-quality and high-quantity samples (Froenicke, 2017). Following sequencing on the HiSeq X platform, we identified a small fraction of reads (<1%) with a wrong combination of barcodes and indices. After excluding several possible explanations, we conclude that index hopping likely happens in this system, but results in a similar rate of read misassignment as reported for older versions of Illumina sequencing platforms. After demonstrating how the use of inline barcode-containing sequencing adapters enables detection and removal of falsely indexed reads, we recommend using this approach independent of the sequencing platform in studies that rely on low-coverage data, require absolute certainty and/or aim to characterize rare variants.

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 Rohland and Reich (2012) and Rohland et al. (2015), as detailed below. 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 (final concentrations: 1× buffer Tango, 100 µM each dNTP, 1 mM ATP, 25 U T4 polynucleotide kinase (Thermo Scientific) 3U USER enzyme (NEB)). 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 (Fig. 1). DNA fragment 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. 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 error (Rohland et al. 2015, Table S1). To increase the complexity of the pooled sequencing library, one sample received two different barcode combinations (Table 1). Adapter ligation was performed in 40 µl volume using 20 µl of blunted DNA and 1 µl of unique P5 and P7 barcodes per sample (final concentrations: 1× T4 DNA ligase buffer, 5% PEG-4000, 5 U T4 DNA ligase (Thermo Scientific), Fig. 1). Samples were incubated for 30 minutes at room temperature and cleaned using MinElute spin columns following the manufacturer's protocol. Adapter fill-in was performed in 40 µl final volume using 20 µl adapter ligated DNA (final concentrations: 1× T4 DNA ligase buffer, 5% PEG-4000, 5 U T4 DNA ligase (Thermo Scientific), Fig. 1), incubated at 37°C for 20 minutes, heat-inactivated at 80°C for 20 minutes, 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 in Meyer & Kircher (2010) (final concentrations: 1x AccuPrime reaction mix, 0.3µM IS4 primer, 0.3µM P7 indexing primer, 7 U AccuPrime Pfx (Thermo Scientific), cycling protocol: 95°C for 2 min, 30 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 1 min and a final extension at 72°C for 5 minutes, Fig. 1). Note that indexing PCR for sample 7 that received two different barcode pairs was performed in a single reaction combining both fractions of this sample. All index sequences differed by at least three base pairs from each other (Table S1). 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 is part of the Illumina sequencing adapter (Fig. 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 library on a Bioanalyzer (Fig. S1). The pooled library with final concentration of 18mM was sequenced on three HiSeq X lanes (150 bp paired-end, 1% PhiX) that were part of independent runs, at the SciLife sequencing facility in Stockholm.

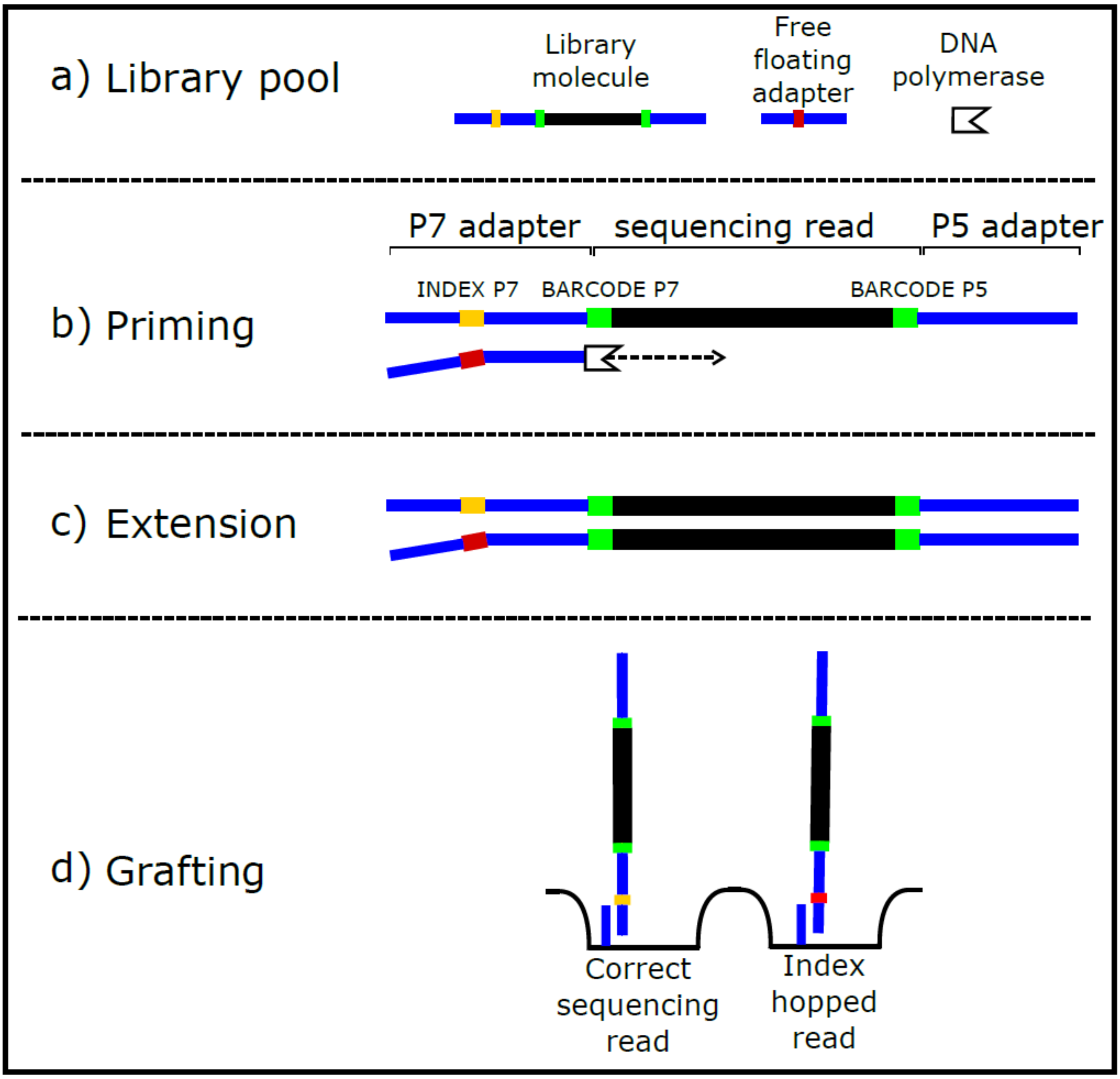


Figure 1: Outcome of index hopping. A) The library pool, containing barcoded and indexed library molecules and free-floating indexing primers, is mixed with ExAmp reagents before loading on the patterned flow cell. B) Free-floating adapters anneal to the adapter sequence of a library molecule and C)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

Data processing

All reads were demultiplexed based on their unique indices using Illumina's bcl2fastq (v2.17.1) software with defaults settings, allowing for one mismatch per index and only retaining "pass filter" reads (Illumina Inc.). All unidentified reads, i.e. reads with indices that were not used in our experiment, were subjected to the same filtering steps, as described below. We removed adapter sequences using AdapterRemoval V2.1.7 using standard parameters and subsequently merged the reads, requiring a minimal overlap of 11bp and allowing for a 10% sequencing error rate (Schubert et al. 2016). Unmerged reads and reads below 29 bp were removed leaving only merged reads with an original insert size of at least 15 bp (7 bp barcodeP7 + 7 bp barcodeP5 + 15 bp DNA fragment = 29 bp). To increase certainty, we only retained reads with intact and error-free P5 and P7 barcodes (assessed using an in-house python script) and an average quality score of at least 30 using prinseq V0.20.4 (Schmieder and Edwards 2011).

Estimating barcode and index cross-contamination and index hopping across sequencing runs

To estimate the rate of barcode cross-contamination, we identified reads with wrong barcode pairs for each sample within each run. We also included unidentified reads with wrong barcode pair combinations into this calculation. The proportion of cross-contaminated reads within a given sequencing run was determined as the ratio between the sum of all reads with wrong barcode pairs and the sum of all sequenced reads that passed the filtering criteria. Given that we used a total of eight different barcodes, we calculated the probability that barcode cross-contamination results in a valid barcode pair (i.e. barcode pair that is actually used in the experiment) as 7*(x/7 * x/7), where x corresponds to the estimated percentage of wrong barcode pairs present in our experiment.

Reads with a correct barcode combination but wrong index can result from index cross-contamination and/or index hopping. To distinguish between these two possibilities, we relied on the fact that only seven different indices were used in our experiment, whereas 40 different indices are routinely used in the ancient DNA laboratory. Therefore, we quantified index cross-contamination as the fraction of reads containing indices that were not included in our experiment. These reads are present within the unidentified reads and carry a valid barcode combination but an unused index.

To determine the proportion of hopped reads, for each sequencing run we calculated the ratio between the sum of all reads showing a wrong index-barcode combination and the sum of all sequenced reads that passed the filtering criteria. To account for the possibility of barcode cross-contamination that produces valid barcode combinations and index cross-contamination, we subtracted these two estimates from the proportion of reads with wrong barcode-index combination.

Statistical analyses

Statistical analyses were performed in R 2.15.3 (Team R Core 2016). Significant global chi-square tests were followed by a post hoc procedure as implemented in the R package polytomous (https://artax.karlin.mff.cuni.cz/r-help/ library/polytomous/html/00Index.html). The minimum value of the chi-squared test statistic for the given degrees of freedom was used to assess if individual observed values differ significantly from an overall hypothetical homogeneous distribution. The test also identified the direction of these differences.

Results

Our sequencing libraries were made from degraded historical samples containing a large proportion of short DNA fragments (Fig. 2A), the majority of which could be confidently merged (95.3% SE ± 1.0%). After filtering (see Methods), the final dataset contained 89.3% ±1.9% of the original sequence reads.

Barcode cross-contamination

We observed low levels of barcode cross-contamination (0.0276% SE ± 0.0026 across all three runs, Table 1, Table S2). 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 (Figure 3, 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 3) is more likely the result of jumping PCR during the 10 rounds of indexing PCR, as both fraction 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. Assuming that adapter ligation of barcodes is unbiased with respect to the barcode sequence (Rohland et al. 2015), the detected low average percentage of cross-contamination will lead to $1.55x10^{-5}$ % of reads (7*(0.00276/7 * 0.00276/7)) * 100% = 0.0000155%) with a valid barcode pair, but wrongly appear as having undergone index hopping.

Index cross-contamination

The Illumina HiSeq X platform does not support a double-indexing design. Therefore, in contrast to barcode cross-contamination, index cross-contamination cannot be directly quantified from the sequencing data. Instead, we focused on the fraction of unidentified reads, which contain indices that were not used in our experiment (Methods, Table S3). The fraction of such reads was nearly identical among the three sequencing runs, ranging from 0.12% to 0.13% (mean = 0.124% SE ± 0.0023).

Index hopping

Index hopping will not affect the barcodes that are directly attached to the DNA fragments. Therefore, it can be readily distinguished from barcode cross-contamination by the presence of reads containing a wrong combination between an index and a barcode pair. Across all three sequencing runs, we detected a low proportion of reads with wrong index-barcode combinations (mean=0.594%, SE ± 0.0434%, Table 1). As detailed in Methods, to obtain the proportion of reads that result from index hopping, but not from barcode or index cross-contamination, we subtracted our estimates of barcode cross-contamination and index cross-contamination from this value. The estimated rate of index hopping in our experiment across all three sequencing runs is therefore 0.470% SE ± 0.044 (0.594% - 1.55×10^{-5} % - 0.124 %). The proportion of hopped reads differed significantly by sample (chi-square test, $P<10^{-15}$). We observed a significant positive correlation between the number of sequenced reads per sample and the number of reads that hopped from this sample to other samples (Pearson's r = 0.96, P = 0.0005), suggesting that samples with higher number of sequenced reads will serve as a dominant source of hopped reads (Fig. 3). Therefore, even though the overall rate of index hopping is low, samples with low number of sequenced reads are more affected by index hopping, leading to 1.47% SE ± 0.11% and 2.49% SE ± 0.29% of index hopped reads within these samples in our experiment (e.g. samples 2 and 4 in Table 1, Table S4, Fig. 3).

The rate of index hopping differed significantly by read length and GC content (chi-square test, $P<10^{-15}$, Figure 2). Reads shorter than 90 bp and reads with GC content above 40% showed significantly higher proportion of hopped reads than expected.

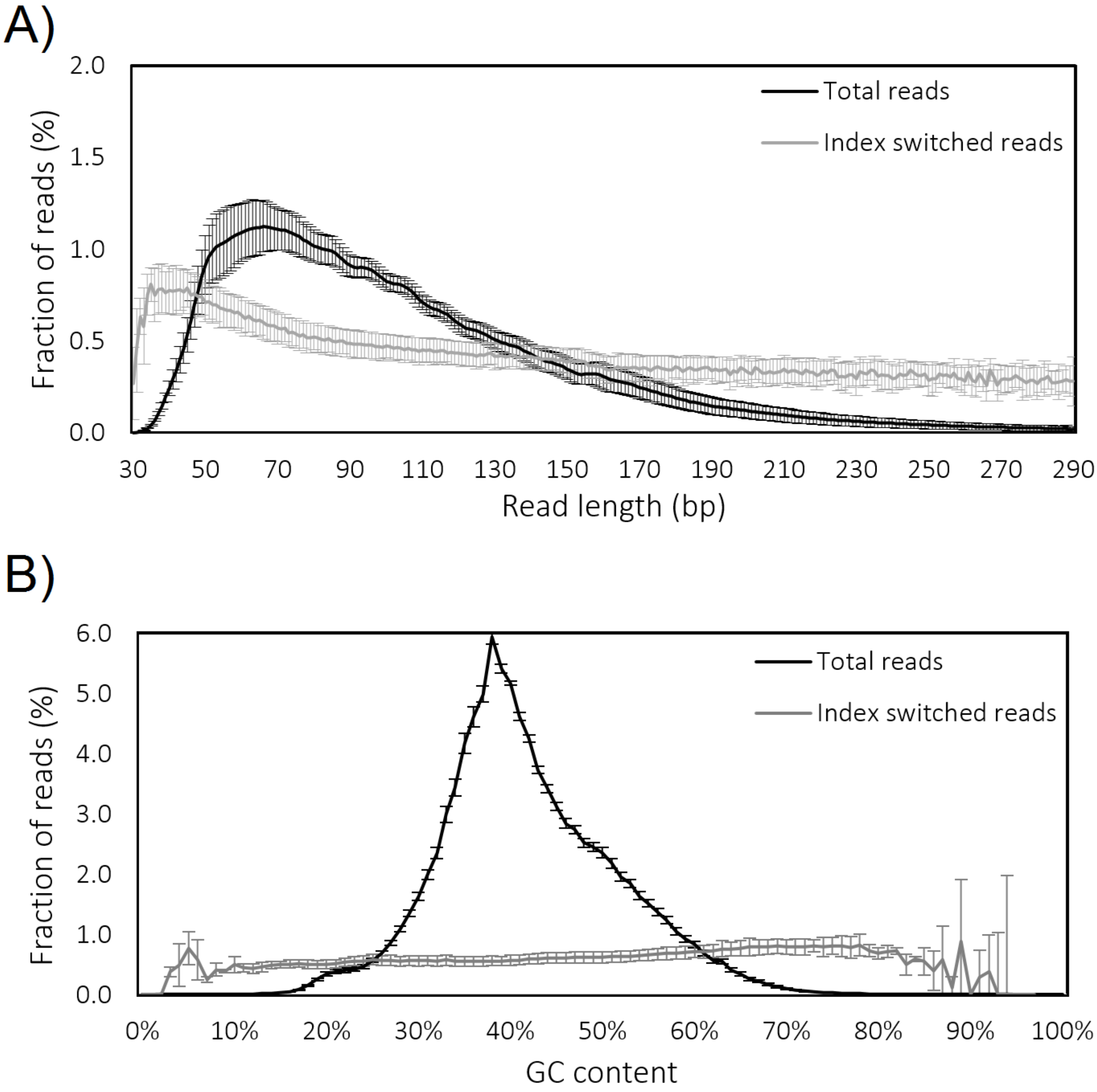


Figure 2: A) Read length distribution and the proportion of index hopping by read length. B) Read GC-content distribution and the proportion of index hopping by read GC content. Vertical bars depict 95% confidence intervals.

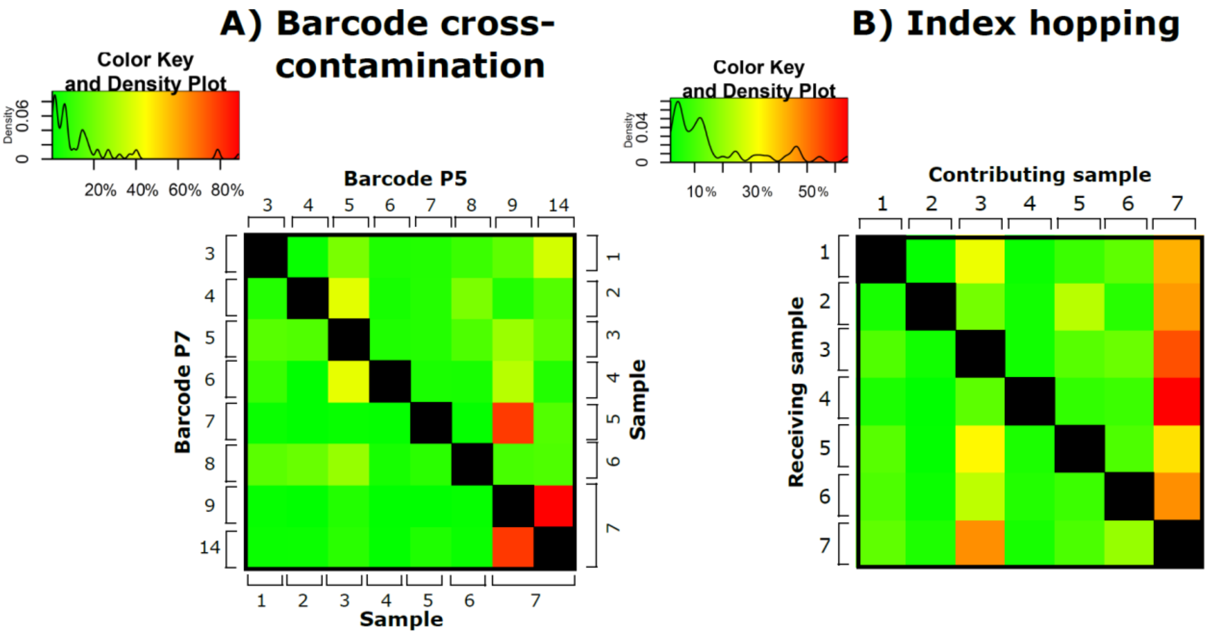


Figure 3: Barcode cross contamination and index hopping by sample. A) Proportion of a given wrong barcode pair in the data out of all erroneous barcode pairs. Barcodes 9 and 14 are paired significantly more often and at equal frequencies, which is likely explained by jumping PCR. B) Proportions of hopped reads by sample. Samples in the top row contribute hopped reads, whereas samples on the left receive 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	9	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 is a real phenomenon occurring on the Illumina HiSeq X platform, but its rate is below 1% in our study. Multiple sources of error can result in read misassignment on the HiSeq X platform, including barcode and index cross-contamination, jumping PCR, sequencing errors, and index hopping. However, through a careful experimental design, we can exclude these error sources and reliably quantify the rate of index hopping. First, we show that the rate of cross-contamination of barcodes is very low (on average, only 0.0027%). A slightly higher level of observed barcode cross-contamination in sample 7 is likely due to jumping-PCR. However, jumping PCR can be eliminated as explanation for wrong index-barcode combinations, as we prepared all libraries individually and avoided amplification of pooled libraries from different samples. Library pooling only occurred directly prior to sequencing. Second, we detect low levels of index cross-contamination by quantifying the presence of indices that are routinely used in the lab among our sequenced reads (0.124%). This further suggests that the presence of wrong index-barcode pairs cannot be explained by index cross-contamination. Third, we employed a very stringent procedure to control for sequencing error: we did not allow for mismatches in the 7-bp P5 and P7 barcodes, required high average read quality and only retained merged reads. By using the library preparation protocol as described in Rohland et al. 2015, we can thus accurately identify and quantify reads containing wrong index-barcode combinations that are the result of index hopping and not the effect of other sources of error.

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 (Kircher et al. 2012; Wright and Vetsigian 2016a, Wright and Vetsigian 2016b) and from 0.06% to 0.21% for the MiSeq platforms (Nelson et al. 2014; 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 the recent estimates for the Illumina HiSeq X/4000 platforms (Griffiths et al. 2017; Owens et al. 2017; Sinha et al. 2017). While (Owens et al. 2017) failed to detect any index hopping in libraries sequenced both on Illumina HiSeq X and HiSeq 2500, (Griffiths et al. 2017) and (Sinha et al. 2017) reported >1% and up to 10% of misassigned reads for single-cell RNA libraries on the HiSeq 4000 platform. Our low observed rate of index hopping might be explained by the low amounts of free-floating adapters during library preparation, since these had been rigorously removed through size selection and cleaning (Figure S1).

The number of reads with hopped indices is proportional to the total number of reads contributed by a given sample to the pooled sequencing library. Pooling samples in unequal amounts 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 displayed up to 3.2% of misassigned reads (Table 1). When working with low-quality samples, the effect of unequal amounts of index hopping can become even more severe if the endogenous content is markedly different between samples, as is often observed in aDNA studies (Damgaard et al. 2015; Pinhasi et al. 2015; van der Valk et al. 2017). In this case, hopping of endogenous reads will occur from samples with high endogenous content into samples with low endogenous content, potentially leading to pronounced biases. The interplay between endogenous content and the number of sequenced reads may result in libraries, in which the proportion of false assigned endogenous reads is considerably higher than reported here (Fig. S2).

Our study shows that while index hopping occurs on the Illumina HiSeq X platform, it results in low proportion of erroneous reads. Importantly, these reads can be readily identified using a library preparation protocol that combines two separate inline-barcodes and a unique index (or index pair on the HiSeq 4000). For studies generating high coverage data, the low detected rate of read misassignment, which is similar to that of the older sequence platforms, might be insignificant. However, in cases where low coverage data is generated or absolutely certainty is required, even low-rate index hopping might represent a major problem. Using short barcode adapters allows for the filtering of misassigned reads, and in the case of short read lengths (such as in aDNA studies) will lead to only a minimal loss of sequencing data. We therefore recommend the use the 7-bp barcode adapters when preparing pooled ancient DNA libraries or in studies were absolute certainty is required.

Acknowledgments:

Funding sources: This project was supported by FORMAS grant 2015-676 to LD and the Jan Löfqvist and the Nilsson-Ehle Endowments of the Royal Physiographic Society of Lund to KG. Sequencing consumables were supplied by Illumina. Illumina had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We also 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.

References

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* **522**: 167–172.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–1624.

Craig DW, Pearson J V, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* **5**: 887–893.

D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**: 55.

Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. 2015. Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep* **5**: 11184.

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

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R. 2013. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**: 979–992.

Froenicke L. 2017. Update on barcode mis-assignment issue. http://dnatech.genomecenter.ucdavis.edu/2017/04/11/update-on-barcode-mis-assignment-issue/

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science (80-)* **328**: 710–722.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446**: 153–8.

Griffiths JA, Lun ATL, Richard AC, Bach K, Marioni JC. 2017. Detection and removal of barcode swapping in single-cell RNA-seq data. *bioRxiv*.

Holcomb CL, Rastrou M, Williams TC, Goodridge D, Lazaro AM, Tilanus M, Erlich HA. 2014. Next-generation sequencing can reveal in vitro-generated PCR crossover products: Some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens* **83**: 32–40.

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

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**: e3.

Lahr DJG, Katz LA. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**: 857–66.

McDevitt SL, Bredeson J V., Roy SW, Lane JA, Noble JA. 2016. HAPCAD: An open-source tool to detect PCR crossovers in next-generation sequencing generated HLA data. *Hum Immunol* **77**: 257–263.

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

Meyerhans A, Vartanian JP, Wain-Hobson S. 1990. DNA recombination during PCR. *Nucleic Acids Res* **18**: 1687–1691.

Mulligan CJ. 2005. Isolation and Analysis of DNA from Archaeological, Clinical, and Natural History

Specimens. *Methods Enzymol* **395**: 87–103.

Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys ed. M.M. Heimesaat. *PLoS One* **9**: e94249.

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

Odelberg SJ, Weiss RB, Hata A, White R. 1995. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res* **23**: 2049–57.

Owens GL, Todesco M, Drummond EBM, Yeaman S, Rieseberg LH. 2017. A Novel Post Hoc Method For Detecting Index Switching Finds No Evidence For Increased Switching On The Illumina HiSeq X. *bioRxiv*.

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 ed. M.D. Petraglia. *PLoS One* **10**: e0129102.

Research & Markets. 2017. Market Assessment & Forecast Global Next Generation Sequencing: 2017 – 2021, Report, ID: 4109654, SA-BRC

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

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

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

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

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**: 14508–13.

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

Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, et al. 2017. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*.

Smith AM, Heisler LE, St.Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, et al. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* **38**: e142–e142.

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/.

Therkildsen NO, Palumbi SR. 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.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46–53.

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

Wall JD, Schlebusch SA, Alberts SC, Cox LA, Snyder-Mackler N, Nevonen KA, Carbone L, Tung J. 2016. Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Mol Ecol* **25**: 3469–3483.

Wright ES, Vetsigian KH. 2016a. Inhibitory interactions promote frequent bistability among competing bacteria. *Nat Commun* **7**: 11274.

Wright ES, Vetsigian KH. 2016b. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics* **17**: 876.

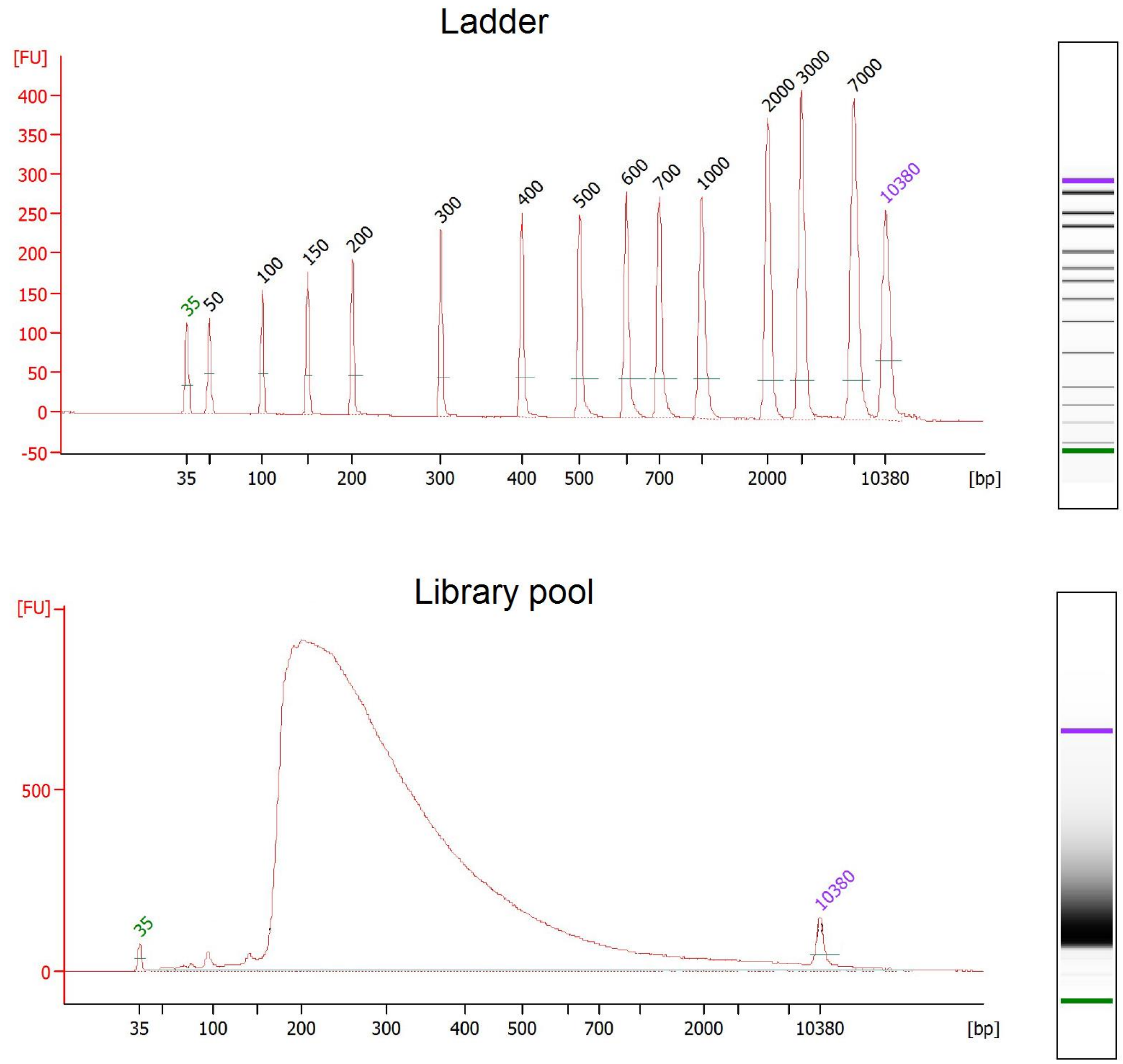


Figure S1: Bioanalyzer profile of the final pooled library. Note that during library preparation, sequencing adapters are attached to the DNA fragments, adding an additional 136 bp to the original DNA fragments. The insert size of the DNA is therefore 136 bp lower than what the Bioanalyzer shows.

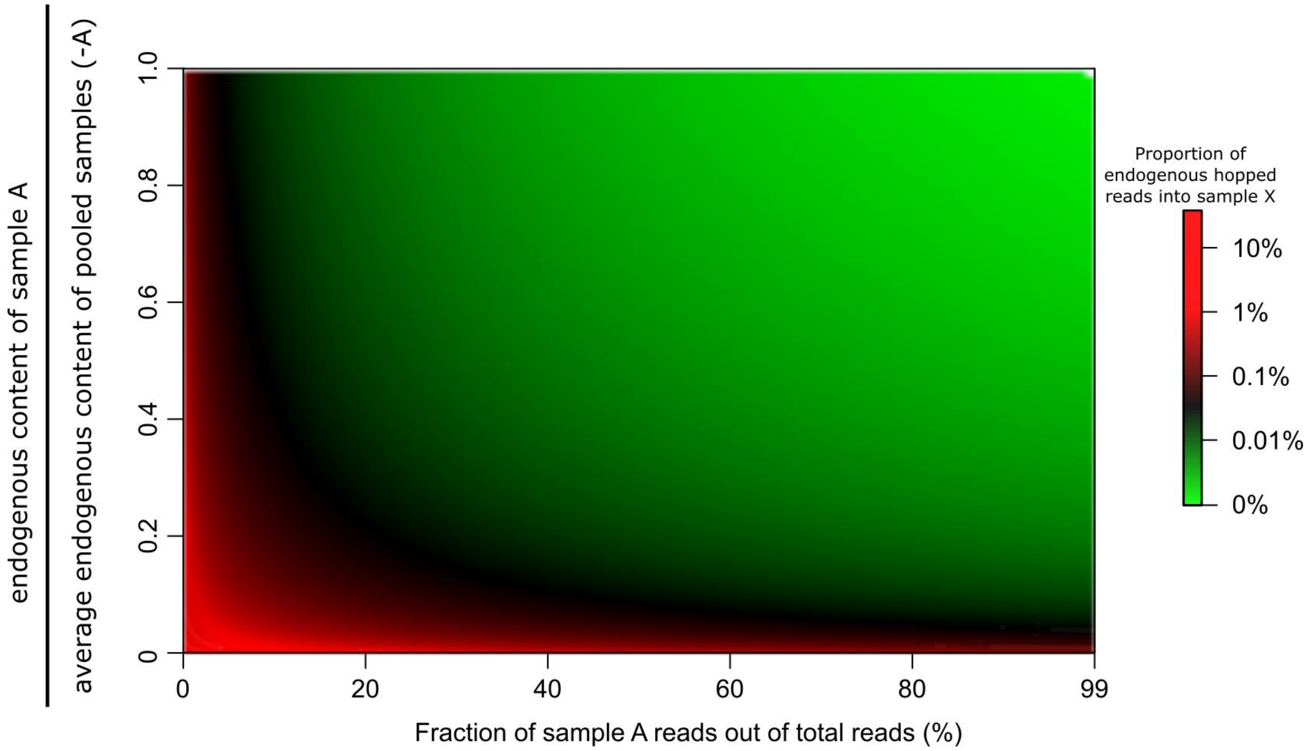
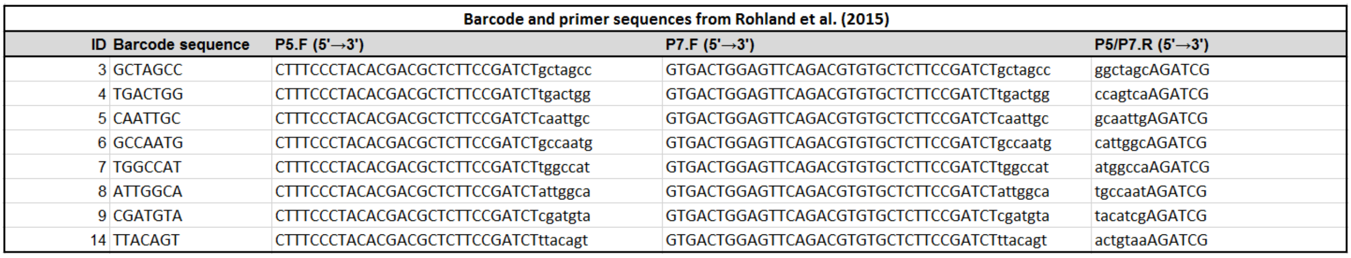
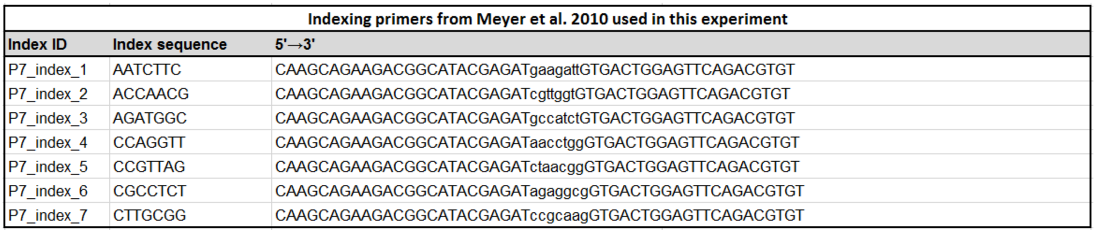


Figure S2. Theoretical relationship between endogenous content, fraction of total reads contributed by a given sample (referred to as sample X) to the pooled sequencing library, and index hopping. The lower the proportion of reads in sample X and the lower its endogenous content compared to other samples in the pooled sequencing library, the higher the proportion of endogenous hopped reads that sample X will receive from other samples.

Table S1, oligonucleotide sequence of the used barcodes and indices



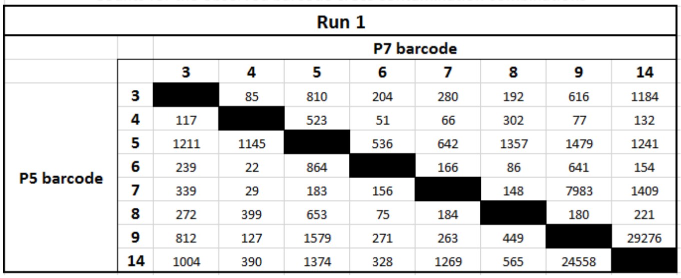
Barcode and primer sequences from Rohland et al. (2015)				
ID	Barcode sequence	P5.F (5'→3')	P7.F (5'→3')	P5/P7.R (5'→3')
3	GCTAGCC	CTTTCCCTACACGACGCTCTTCCGATCTgctagcc	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgctagcc	ggctagcAGATCG
4	TGACTGG	CTTTCCCTACACGACGCTCTTCCGATCTtgactgg	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtgactgg	ccagtcaAGATCG
5	CAATTGC	CTTTCCCTACACGACGCTCTTCCGATCTcaattgc	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcaattgc	gcaattgAGATCG
6	GCCAATG	CTTTCCCTACACGACGCTCTTCCGATCTgccaatg	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgccaatg	cattggcAGATCG
7	TGGCCAT	CTTTCCCTACACGACGCTCTTCCGATCTtggccat	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtggccat	atggccaAGATCG
8	ATTGGCA	CTTTCCCTACACGACGCTCTTCCGATCTattggca	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTattggca	tgccaatAGATCG
9	CGATGTA	CTTTCCCTACACGACGCTCTTCCGATCTcgatgta	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcgatgta	tacatcgAGATCG
14	TTACAGT	CTTTCCCTACACGACGCTCTTCCGATCTttacagt	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTttacagt	actgtaaAGATCG



Indexing primers from Meyer et al. 2010 used in this experiment		
Index ID	Index sequence	5'→3'
P7_index_1	AATCTTC	CAAGCAGAAGACGGCATACGAGATgaagattGTGACTGGAGTTCAGACGTGT
P7_index_2	ACCAACG	CAAGCAGAAGACGGCATACGAGATcgttggtGTGACTGGAGTTCAGACGTGT
P7_index_3	AGATGGC	CAAGCAGAAGACGGCATACGAGATgccatctGTGACTGGAGTTCAGACGTGT
P7_index_4	CCAGGTT	CAAGCAGAAGACGGCATACGAGATaacctggGTGACTGGAGTTCAGACGTGT
P7_index_5	CCGTTAG	CAAGCAGAAGACGGCATACGAGATctaacggGTGACTGGAGTTCAGACGTGT
P7_index_6	CGCCTCT	CAAGCAGAAGACGGCATACGAGATagaggcgGTGACTGGAGTTCAGACGTGT
P7_index_7	CTTGCGG	CAAGCAGAAGACGGCATACGAGATccgcaagGTGACTGGAGTTCAGACGTGT

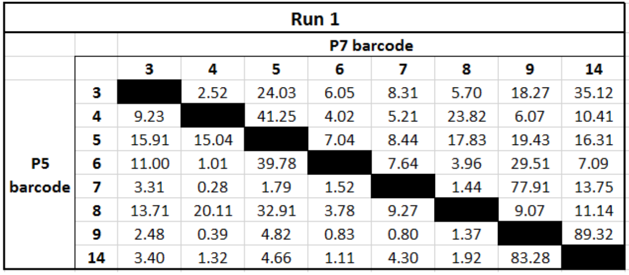
Table S2, Barcode cross-contamination. Sample on the left is the receiving sample, sample on top is the contributing sample.

Counts for the observed barcode cross contamination combinations

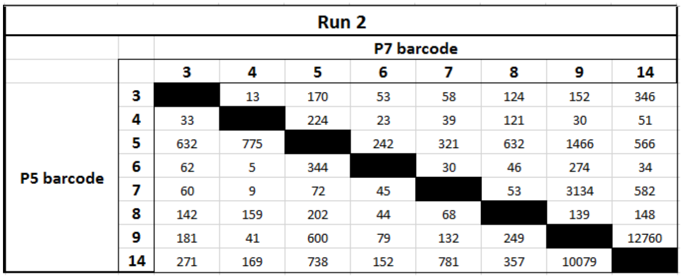


Run 1									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	85	810	204	280	192	616	1184
	4	117	■	523	51	66	302	77	132
	5	1211	1145	■	536	642	1357	1479	1241
	6	239	22	864	■	166	86	641	154
	7	339	29	183	156	■	148	7983	1409
	8	272	399	653	75	184	■	180	221
	9	812	127	1579	271	263	449	■	29276
	14	1004	390	1374	328	1269	565	24558	■

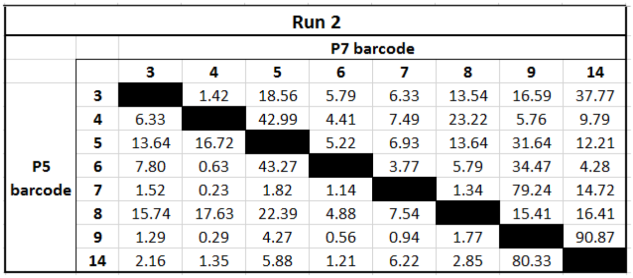
Percentage of reads containing a given barcode combination out of total cross-contaminated reads containing this barcode



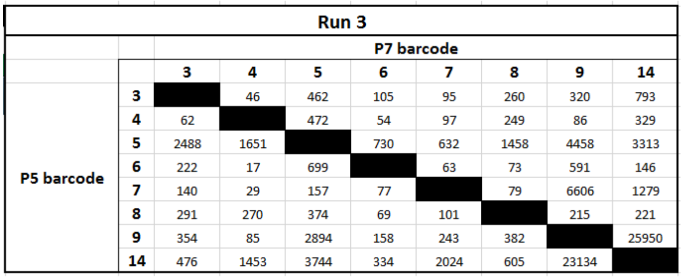
Run 1									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	2.52	24.03	6.05	8.31	5.70	18.27	35.12
	4	9.23	■	41.25	4.02	5.21	23.82	6.07	10.41
	5	15.91	15.04	■	7.04	8.44	17.83	19.43	16.31
	6	11.00	1.01	39.78	■	7.64	3.96	29.51	7.09
	7	3.31	0.28	1.79	1.52	■	1.44	77.91	13.75
	8	13.71	20.11	32.91	3.78	9.27	■	9.07	11.14
	9	2.48	0.39	4.82	0.83	0.80	1.37	■	89.32
	14	3.40	1.32	4.66	1.11	4.30	1.92	83.28	■



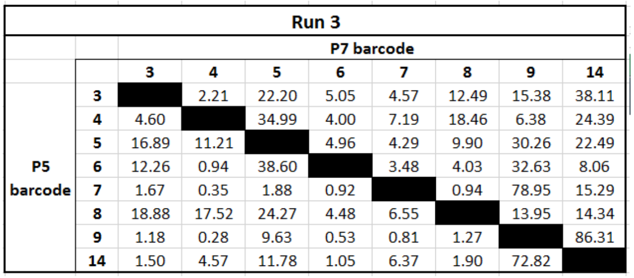
Run 2									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	13	170	53	58	124	152	346
	4	33	■	224	23	39	121	30	51
	5	632	775	■	242	321	632	1466	566
	6	62	5	344	■	30	46	274	34
	7	60	9	72	45	■	53	3134	582
	8	142	159	202	44	68	■	139	148
	9	181	41	600	79	132	249	■	12760
	14	271	169	738	152	781	357	10079	■



Run 2									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	1.42	18.56	5.79	6.33	13.54	16.59	37.77
	4	6.33	■	42.99	4.41	7.49	23.22	5.76	9.79
	5	13.64	16.72	■	5.22	6.93	13.64	31.64	12.21
	6	7.80	0.63	43.27	■	3.77	5.79	34.47	4.28
	7	1.52	0.23	1.82	1.14	■	1.34	79.24	14.72
	8	15.74	17.63	22.39	4.88	7.54	■	15.41	16.41
	9	1.29	0.29	4.27	0.56	0.94	1.77	■	90.87
	14	2.16	1.35	5.88	1.21	6.22	2.85	80.33	■



Run 3									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	46	462	105	95	260	320	793
	4	62	■	472	54	97	249	86	329
	5	2488	1651	■	730	632	1458	4458	3313
	6	222	17	699	■	63	73	591	146
	7	140	29	157	77	■	79	6606	1279
	8	291	270	374	69	101	■	215	221
	9	354	85	2894	158	243	382	■	25950
	14	476	1453	3744	334	2024	605	23134	■



Run 3									
		P7 barcode							
		3	4	5	6	7	8	9	14
P5 barcode	3	■	2.21	22.20	5.05	4.57	12.49	15.38	38.11
	4	4.60	■	34.99	4.00	7.19	18.46	6.38	24.39
	5	16.89	11.21	■	4.96	4.29	9.90	30.26	22.49
	6	12.26	0.94	38.60	■	3.48	4.03	32.63	8.06
	7	1.67	0.35	1.88	0.92	■	0.94	78.95	15.29
	8	18.88	17.52	24.27	4.48	6.55	■	13.95	14.34
	9	1.18	0.28	9.63	0.53	0.81	1.27	■	86.31
	14	1.50	4.57	11.78	1.05	6.37	1.90	72.82	■

Table S3, Indexing primers from Meyer et al. 2010 which are handled in the aDNA lab but not used in this study and the number of reads containing the respective index per run

Index ID	Oligo sequence (5'-3') (index marked with small letters)	Index sequence (5'-3')	Reads with respective index		
			Run 1	Run 2	Run 3
P7_index_8	CAAGCAGAAGACGGCATACGAGATttcgagcGTGACTGGAGTTCAGACGTGT	GCTCGAA	5348	1931	7126
P7_index_9	CAAGCAGAAGACGGCATACGAGATagttggtGTGACTGGAGTTCAGACGTGT	ACCAACT	112581	47853	149981
P7_index_10	CAAGCAGAAGACGGCATACGAGATgtaccggGTGACTGGAGTTCAGACGTGT	CCGGTAC	4733	1861	6352
P7_index_11	CAAGCAGAAGACGGCATACGAGATcggagttGTGACTGGAGTTCAGACGTGT	AACTCCG	1521	586	2081
P7_index_12	CAAGCAGAAGACGGCATACGAGATacttcaaGTGACTGGAGTTCAGACGTGT	TTGAAGT	10938	4811	14492
P7_index_13	CAAGCAGAAGACGGCATACGAGATtgatagtGTGACTGGAGTTCAGACGTGT	ACTATCA	1860	751	2672
P7_index_14	CAAGCAGAAGACGGCATACGAGATgatccaaGTGACTGGAGTTCAGACGTGT	TTGGATC	3734	1673	5052
P7_index_15	CAAGCAGAAGACGGCATACGAGATcaggtcgGTGACTGGAGTTCAGACGTGT	CGACCTG	7975	3037	10605
P7_index_16	CAAGCAGAAGACGGCATACGAGATcgcattaGTGACTGGAGTTCAGACGTGT	TAATGCG	20306	8057	29426
P7_index_17	CAAGCAGAAGACGGCATACGAGATggtacctGTGACTGGAGTTCAGACGTGT	AGGTACC	9312	3633	12493
P7_index_18	CAAGCAGAAGACGGCATACGAGATggacgcaGTGACTGGAGTTCAGACGTGT	TGCGTCC	19341	7288	25033
P7_index_19	CAAGCAGAAGACGGCATACGAGATgagattcGTGACTGGAGTTCAGACGTGT	GAATCTC	3421	1068	4663
P7_index_20	CAAGCAGAAGACGGCATACGAGATgagcatgGTGACTGGAGTTCAGACGTGT	CATGCTC	157222	53466	216544
P7_index_21	CAAGCAGAAGACGGCATACGAGATgttgcgtGTGACTGGAGTTCAGACGTGT	ACGCAAC	11657	4692	14907
P7_index_22	CAAGCAGAAGACGGCATACGAGATccaatgcGTGACTGGAGTTCAGACGTGT	GCATTGG	28894	10347	42820
P7_index_23	CAAGCAGAAGACGGCATACGAGATcgagatcGTGACTGGAGTTCAGACGTGT	GATCTCG	464	152	901
P7_index_24	CAAGCAGAAGACGGCATACGAGATcatattgGTGACTGGAGTTCAGACGTGT	CAATATG	1	2	13
P7_index_25	CAAGCAGAAGACGGCATACGAGATgacgtcaGTGACTGGAGTTCAGACGTGT	TGACGTC	5	1	8
P7_index_26	CAAGCAGAAGACGGCATACGAGATtggcatcGTGACTGGAGTTCAGACGTGT	GATGCCA	22	3	29
P7_index_27	CAAGCAGAAGACGGCATACGAGATgtaattgGTGACTGGAGTTCAGACGTGT	CAATTAC	35	14	284
P7_index_28	CAAGCAGAAGACGGCATACGAGATcctatctGTGACTGGAGTTCAGACGTGT	AGATAGG	1	0	5
P7_index_29	CAAGCAGAAGACGGCATACGAGATcaatcggGTGACTGGAGTTCAGACGTGT	CCGATTG	354	90	244
P7_index_30	CAAGCAGAAGACGGCATACGAGATgcggcatGTGACTGGAGTTCAGACGTGT	ATGCCGC	31	39	331
P7_index_31	CAAGCAGAAGACGGCATACGAGATagtactgGTGACTGGAGTTCAGACGTGT	CAGTACT	0	0	0
P7_index_32	CAAGCAGAAGACGGCATACGAGATtactattGTGACTGGAGTTCAGACGTGT	AATAGTA	56	21	11
P7_index_193	CAAGCAGAAGACGGCATACGAGATcaacctcGTGACTGGAGTTCAGACGTGT	GAGGTTG	137	6	51
P7_index_194	CAAGCAGAAGACGGCATACGAGATactcataGTGACTGGAGTTCAGACGTGT	TATGAGT	7	2	3
P7_index_195	CAAGCAGAAGACGGCATACGAGATaacgaagGTGACTGGAGTTCAGACGTGT	CTTCGTT	798	213	151
P7_index_196	CAAGCAGAAGACGGCATACGAGATcggacggGTGACTGGAGTTCAGACGTGT	CCGTCCG	12	4	22
P7_index_197	CAAGCAGAAGACGGCATACGAGATtaacgttGTGACTGGAGTTCAGACGTGT	AACGTTA	73	20	62
P7_index_198	CAAGCAGAAGACGGCATACGAGATaatatgcGTGACTGGAGTTCAGACGTGT	GCATATT	2	4	9
P7_index_199	CAAGCAGAAGACGGCATACGAGATggaaggtGTGACTGGAGTTCAGACGTGT	ACCTTCC	26	9	282
P7_index_200	CAAGCAGAAGACGGCATACGAGATctcggaaGTGACTGGAGTTCAGACGTGT	TTCCGAG	27	16	8

Table S4, Index hopped reads. Sample on the left is the receiving sample, sample on top is the contributing sample.

Index hopped reads (absolute read count)

Run 1							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		2727	49532	4381	12904	18797	70564
Sample 2	5088		19135	5173	31311	7378	62040
Sample 3	29090	9081		9348	32066	37179	148583
Sample 4	11292	2993	30318		18709	21484	177968
Sample 5	20233	2826	52297	8077		16747	66058
Sample 6	8383	2141	18866	4888	6769		37380
Sample 7	64580	23698	230951	21080	48316	92810	

Run 2							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		1191	23399	2386	8643	13576	38502
Sample 2	3098		12342	2503	16405	5424	37730
Sample 3	14989	4392		4720	22662	21745	94994
Sample 4	6572	1759	16249		11376	12195	79266
Sample 5	8679	1185	31446	4569		9975	30554
Sample 6	4205	866	7019	2120	3831		20965
Sample 7	31484	11790	115066	10011	29857	58186	

Run 3							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		8536	96273	9467	22791	33267	110474
Sample 2	9502		31538	7841	51120	12393	74118
Sample 3	79602	29426		26470	62995	93300	280994
Sample 4	17495	5672	50783		24483	30048	220642
Sample 5	44772	8938	100035	14343		34682	105394
Sample 6	18443	6711	52288	8974	13100		61840
Sample 7	126548	49226	414703	41078	91523	158784	

Percentage of hopped reads out of total hopped reads

Run 1							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		6.27	12.35	8.27	8.60	9.67	12.54
Sample 2	3.67		4.77	9.77	20.86	3.80	11.03
Sample 3	20.98	20.89		17.66	21.37	19.13	26.41
Sample 4	8.14	6.89	7.56		12.47	11.05	31.63
Sample 5	14.59	6.50	13.04	15.25		8.61	11.74
Sample 6	6.05	4.93	4.70	9.23	4.51		6.64
Sample 7	46.57	54.52	57.58	39.81	32.19	47.74	

Run 2							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		5.62	11.39	9.07	9.32	11.21	12.75
Sample 2	4.49		6.01	9.51	17.68	4.48	12.49
Sample 3	21.71	20.73		17.94	24.43	17.96	31.45
Sample 4	9.52	8.30	7.91		12.26	10.07	26.25
Sample 5	12.57	5.59	15.30	17.37		8.24	10.12
Sample 6	6.09	4.09	3.42	8.06	4.13		6.94
Sample 7	45.61	55.66	55.99	38.05	32.18	48.05	

Run 3							
	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
Sample 1		7.87	12.91	8.75	8.57	9.18	8.46
Sample 2	3.21		4.23	7.25	19.22	3.42	5.67
Sample 3	26.86	27.12		24.47	23.68	25.74	21.51
Sample 4	5.90	5.23	6.81		9.20	8.29	16.89
Sample 5	15.11	8.24	13.42	13.26		9.57	8.07
Sample 6	6.22	6.18	7.01	8.30	4.92		4.73
Sample 7	42.70	45.37	55.62	37.97	34.41	43.81	