

Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform

Martin Kircher*, Susanna Sawyer and Matthias Meyer*

Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics,
04103 Leipzig, Germany

Received July 6, 2011; Revised August 9, 2011; Accepted September 3, 2011

ABSTRACT

Due to the increasing throughput of current DNA sequencing instruments, sample multiplexing is necessary for making economical use of available sequencing capacities. A widely used multiplexing strategy for the Illumina Genome Analyzer utilizes sample-specific indexes, which are embedded in one of the library adapters. However, this and similar multiplex approaches come with a risk of sample misidentification. By introducing indexes into both library adapters (double indexing), we have developed a method that reveals the rate of sample misidentification within current multiplex sequencing experiments. With ~0.3% these rates are orders of magnitude higher than expected and may severely confound applications in cancer genomics and other fields requiring accurate detection of rare variants. We identified the occurrence of mixed clusters on the flow as the predominant source of error. The accuracy of sample identification is further impaired if indexed oligonucleotides are cross-contaminated or if indexed libraries are amplified in bulk. Double-indexing eliminates these problems and increases both the scope and accuracy of multiplex sequencing on the Illumina platform.

INTRODUCTION

Over the last decade, new sequencing technologies have become available (1–5), which greatly outperform the older Sanger technology in terms of throughput and cost. Due to the large number of sequences generated with these technologies, there is a growing interest in sequencing multiple samples in parallel. Using different library construction protocols, sample-specific index sequences (also called ‘barcodes’) can be attached to the sample molecules during sequencing library preparation

[e.g. (6–9)]. Subsequently, multiple libraries can be pooled and sequenced in the same region, and later computationally separated based on their index sequence. This facilitates highly parallel sequencing of a large number of samples (96 and more).

On the Illumina sequencing platform, the perhaps most widely used multiplexing strategy (the vendor’s protocol) uses indexes, which are embedded within one of the adapters (8–10), separated from the actual template (Figure 1A). Thus, for a typical Illumina multiplex library, the index is sequenced after the forward read in a separate ‘index read’, for which a new sequencing primer is annealed. Although there are alternative indexing approaches, where the index is attached adjacent to the insert [e.g. (7)], this strategy has several benefits. Decoupling the actual template read and the index read allows the index read to be left out if not required and also keeps the sequencing error rate low, because phasing (3,11,12), one of the main sources of sequencing error on the Illumina platform, is reset with the annealing of a new sequencing primer. In addition, image analysis and the estimation of base calling parameters are not affected by the frequently unbalanced base composition of indexes.

While sample multiplexing greatly increases experimental scalability, it also introduces the danger of falsely assigning sequences to their original samples. Some applications, however, require highly accurate genotyping, particularly if conclusions are drawn from the occurrence of rare sequence variants. These could for example be rare transcripts or somatic mutations. In cancer research, for example, low-frequency somatic mutations can harbor important biological insights (13). The current throughput of Illumina sequencers is sufficiently high for sequencing exomes of several tumor/normal pairs to high coverage. It can therefore be anticipated that multiplex sequencing will soon be a common tool in many biomedical studies. Another very sensitive application—and the initial motivation behind this study—is ancient DNA research. Here, the observation of a single sequence may be taken as evidence for DNA survival or presence of contamination (14,15).

*To whom correspondence should be addressed. Tel: +49 341 3550 500; Fax: +49 341 3550 555; Email: martin.kircher@eva.mpg.de
Correspondence may also be addressed to Matthias Meyer. Tel: +49 341 3550 509; Fax: +49 341 3550 555; Email: mmeyer@eva.mpg.de

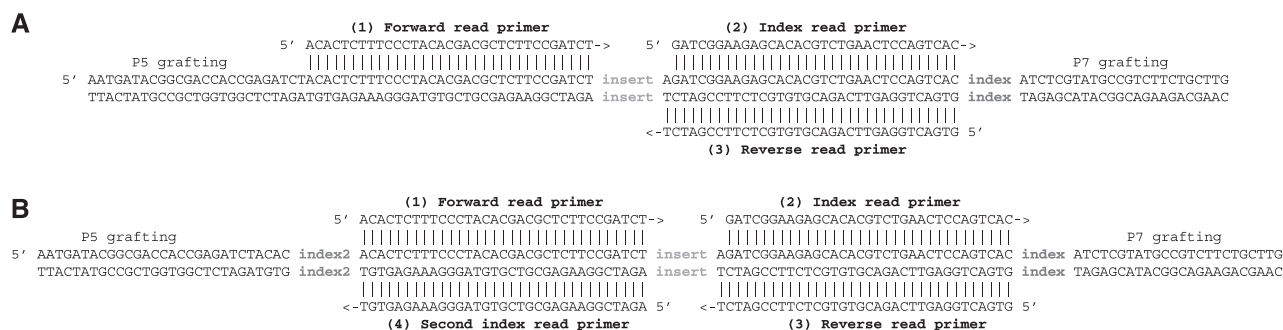


Figure 1. (A) Regular Illumina multiplex library design. The grafting sequences (P5 and P7) are used for template immobilization and amplification. Three distinct sequence reads (forward read, index read, reverse read) are primed from different adapter sites. (B) Double-index library design with an additional index incorporated into the second adapter. Here, four distinct sequence reads are performed.

To avoid false-assignments of sequences to samples, previous studies have focused on generating highly distinguishable index designs (6,8,9,16), i.e. requiring several sequence changes before one index sequence is converted into another valid index sequence. This should efficiently reduce the number of index conversions due to errors in sequencing, library amplification or oligonucleotide synthesis. For example, assuming an error rate of 0.5% per position, an edit distance of three would correspond to a false-assignment probability of 4.31×10^{-6} for 7-nt indexes (~172 per 40 million sequences in a lane of an Illumina Genome Analyzer IIx flow cell) under a simple binomial error distribution model. This is much lower than required for most applications. Another possible source of sample misidentification is cross-contamination of indexes by the oligonucleotide manufacturer or during later handling. This can be caused, for example, by sequential purification of different indexing primer/adaptor oligonucleotides on the same high performance liquid chromatography (HPLC) column after synthesis. Even though columns are washed between oligonucleotides, low levels of carry-over contamination may be difficult to prevent.

In many cases, sequencing libraries need to be enriched for certain targets of interest, often by means of hybridization capture, before they are subjected to multiplex sequencing. To minimize efforts, it can be desirable to pool samples prior to target capture to perform both capture and sequencing in a multiplex setup. Since it is usually inevitable to amplify libraries after capture, this strategy introduces a step where libraries from different samples are amplified in a single reaction. Unfortunately, PCR can produce chimeras by recombining different templates molecules (a process often referred to as 'jumping PCR') (17–20). Consequently, multiplex capture may introduce significant levels of sample cross-contamination.

In order to better quantify and improve the accuracy of multiplexing sequencing on the Illumina platform, we have devised a new double-indexing method, which places indexes into both of the universal adapter sequences (Figure 1B), thereby extending the current system from three to four sequencing reads. Using Neandertal DNA extracts, we constructed double-indexed libraries with unique index combinations and performed three experiments. First, libraries were pooled only for sequencing

(experiment no-CAP). Second, the same libraries were enriched individually for mitochondrial DNA using a recently published hybridization capture method (21) and then pooled for sequencing (experiment SP-CAP). Third, libraries were pooled prior to target enrichment, and hence capture, amplification and sequencing were all performed in a multiplex setup (MP-CAP), allowing for possible cross-contamination, caused by jumping PCR, to be quantified.

MATERIALS AND METHODS

Library preparation, amplification and target enrichment

Fifteen DNA extracts (L1–L15) were prepared using 100–200 mg of bone powder from two ~30 000-year-old Neandertal bones (Vi33.25 and Vi33.26 from Vindija Cave, Croatia) following the protocol of Rohland *et al.* (22). Two negative controls (L16 and L17) were carried through the extraction process. Sequencing libraries were prepared from the extracts using a previously published protocol (9) with the following modifications: (i) All SPRI purification steps were substituted by spin column purification (MinElute PCR purification kit, Qiagen). (ii) For L11–L15 and L17, USER enzyme mix (New England Biolabs) was added to the blunt-end repair reaction to remove uracils (23). (iii) Adapter concentration in the ligation reaction was reduced to 0.25 μ M of each adapter. (iv) No purification step was performed after adapter fill-in with Bst polymerase. Instead, the enzyme was heat inactivated at 80°C for 20 min. The reaction mix was then used directly as template for PCR.

All libraries were amplified twice by PCR, using a polymerase that is capable of copying across deoxyuracils for the first, and a proof-reading polymerase for the second amplification. Using 5'-tailed primers ('indexing primers'; see Supplementary Table S1 for all primer sequences), indexes were added to both ends of the library molecules during the first amplification. The entire library volumes were used as templates in 100 μ l PCR reactions containing 1 $\times$ Thermopol buffer (NEB), 5 U AmpliTaq Gold (Applied Biosystems), 250 μ M each dNTP and 400 nM each indexing primer. Cycling conditions were comprised of an activation step lasting 12 min at 95°C, followed by 10 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 30 s and elongation at 72°C for 40 s, with a

final extension step at 72°C for 5 min. The index combinations used for each library are listed in Supplementary Table S2. PCR products were purified using the MinElute PCR purification kit and eluted in 20 µl EB. An amount of 5 µl of the eluates were used as template for the second round of amplification, which was performed in 100 µl reactions containing 1× Phusion High Fidelity Mastermix (NEB) and the primers IS5 and IS6 (9) at a concentration of 400 nM each. Cycling conditions were comprised of an activation step lasting 30 s at 98°C, followed by 10 cycles of denaturation at 98°C for 20 s, annealing at 60°C for 30 s and elongation at 72°C for 40 s, with a final extension step at 72°C for 5 min. PCR products were purified using the MinElute PCR purification kit and eluted in 10 µl EB. The concentrations of all libraries were determined on a Bioanalyzer 2100 (Agilent) using DNA 1000 chips.

Libraries were either directly pooled and sequenced (no-CAP experiment) or enriched for mitochondrial DNA. Enrichment was performed either individually (experiment SP-CAP) or in bulk (experiment MP-CAP) using a protocol detailed in Maricic *et al.* (21). After enrichment, the libraries in the SP-CAP and MP-CAP experiments were amplified for 24 cycles using Phusion polymerase under the conditions described above. Libraries were purified using the MinElute PCR purification kit, quantified on a Bioanalyzer 2100 and pooled in equimolar ratios.

Sequencing

Libraries were sequenced in three lanes of an Illumina Genome Analyzer IIx run (v4 chemistry, v2 cluster generation kit). Deviating from the manufacturer's instruction for a $2 \times 101 + 7$ cycles multiplexed paired end run, a ϕ X174 control library was spiked into all lanes, contributing to on average about 1% of the reads in each lane. Furthermore, an additional seven-cycle index read was performed by repeating the chemistry steps of the first index (without commands marking this part of the read as index) at the end of the run recipe. This second index read used the custom sequencing primer shown in Figure 1B.

Data processing

The sequencing data was analyzed three times, once starting from QSEQ sequence files and CIF intensity files obtained from Illumina's Genome Analyzer SCS 2.6/RTA 1.6 software, and twice starting from raw images using OLB 1.8 and OLB 1.9. In all cases, the QSEQ raw reads obtained from Illumina's base caller Bustard were aligned to the ϕ X174 reference sequence to obtain a training data set for the base caller Ibis 1.1.2 (12), which was then used to call bases and quality scores. The PF flag for each cluster was extracted from the QSEQ files of the Illumina pipeline output. Index pairs were analyzed starting from raw sequences and considering only perfect matches to the index sequences.

Sequences from OLB 1.8 intensity files (Ibis called) have been deposited in the ENA with accession number ERP000829.

RESULTS

Quantifying false index pairings

In contrast to current single-indexed multiplex sequencing, double-indexing allows for determining the sample origin of each sequence twice independently. Considering only perfect matches to the designated index sequences, we first compared the two index sequences to estimate the fraction of sequences with conflicting information on sample origin. From the previous considerations—based on a 0.5% sequencing error rate as the only source of error—we would expect to find approximately 172 false index pairs in 20 million reads (<0.001%). Using the sequencing error rate and other parameters of the actual experiments, we expect even fewer false index pairs (4 in 20 million reads, see Supplementary Methods section). However, in stark contrast to these expectations, we found 0.582% in no-CAP, 0.509% in SP-CAP and 0.760% in MP-CAP of the index pairings to be wrong (Table 1). Interestingly, we observed extremely high fractions of false index pairs irrespective of whether libraries were only sequenced together (no-CAP, SP-CAP) or also amplified together (MP-CAP), indicating that factors other than jumping PCR must contribute substantially to the fraction of false pairings.

Removing mixed clusters with signal purity filters

To elucidate the major source of false index pairings, we first checked whether false pairs accumulate at the edges or in specific regions of the flow cell image tiles, but could not see any spatial pattern when overlaying the X,Y-coordinates for correct and false index pairs (data not shown). We then applied Illumina's Pass Filter (PF) flag to the raw sequence data, which is a widely used filter based on the signal purity of each cluster over the first 25 bases of the sequencing run. This filter is supposed to reduce the number of sequences from mixed clusters (i.e. PCR product colonies derived from more than one template molecule). In our experiments, ~80% of the sequences passed this filter (Table 1). Albeit unsatisfactory, we in fact observed a reduction in the fraction of falsely paired indexes (e.g. from 0.582% to 0.523% in no-CAP), suggesting that mixed clusters could be the source of these falsely paired indexes. To test this hypothesis, we manually checked the raw intensity signals from a few clusters with conflicting index reads. In all cases, we detected overlaying signals from at least two different sequence populations (see Supplementary Figure S1 and Supplementary Methods section). In addition, we analyzed sequence reads with conflicting index information from very short molecules, which are present in libraries constructed from ancient DNA. There, sequencing proceeds through the insert into the adapter, providing yet another independent observation of the index sequences. In almost all cases the template read generated a congruent index pair (311 of 311 in experiment no-CAP, 141 of 149 in SP-CAP, 99 of 109 in MP-CAP; see Supplementary Methods section), providing further evidence for the occurrence of mixed clusters on the flow cell.

Table 1. Numerical summary of the false-assignment rates and fractions of false index pairs observed for the three different experiments no-CAP, SP-CAP and MP-CAP

	no-CAP	SP-CAP	MP-CAP
Total number of raw reads	34 241 955	48 546 372	34 684 183
Index pairs in raw data			
Correct pairs (*) (%)	89.14	78.83	89.38
False pairs (%)	0.582	0.509	0.760
False index pairs after PF-filtering of raw reads			
Total number of PF-filtered reads	27 466 817	37 586 292	27 220 161
False pairs (%)	0.523	0.387	0.691
False index pairs after quality score based filtering of index reads			
Average index quality filter (~PF) (%)	0.059	0.192	0.423
Minimum index quality filter (~PF) (%)	0.060	0.177	0.422
Minimum index quality score of 15 (%)	0.060	0.138	0.428
False index pairs after quality score based filtering of template reads			
Read quality filter on both reads (~PF) (%)	0.362	0.439	0.614
Read quality filter on the forward read (~PF) (%)	0.389	0.394	0.593
Read quality filter on the reverse read (~PF) (%)	0.298	—	—
Quantifying cross-contamination, mixed clusters and jumping PCR			
False pairs due to contamination (%)	0.042	0.104	0.038
False pairs due to mixed clusters / jump. PCR (%)	0.018	0.034	0.390

~PF indicates values for a quality score cutoff that removes fewer raw reads than Illumina’s Pass Filter (PF) flag.
*The fraction of correct index pairs is strongly affected by loading density. Denser loading in experiment SP-CAP led to a higher sequencing error rate and hence reduced the fraction of correct index pairs.

To efficiently eliminate sequences from mixed clusters, we explored the effect of applying a base quality score filter specifically to the index reads. Base quality scores are highly correlated with signal purity, but also incorporate signal strength. We considered a filter based on the average base quality score across the two index reads and another filter based on the minimum base quality score observed in the index reads. Using cut-offs that remove just a little less raw data than the Pass Filter flag, both filters remove considerably more false pairs than the Pass Filter flag (e.g. 0.059%/0.060% versus 0.523% false pairs remaining in no-CAP; Table 1). While all three experiments show similar trajectories for the different filter cut-offs (Figure 2), we note that the fraction of false pairs is always much higher for the MP-CAP experiment and that higher cluster densities in the SP-CAP experiment seem to negatively affect quality scores, as more data is removed using the same score cut-off (dashed black lines, Figure 2). To check whether the quality scores in the forward and reverse template reads also correlate with the fraction of false index pairs, we applied a minimum quality filter on those reads as well (Table 1 and Supplementary Figure S2). Although this filter is also more efficient than the PF flag, it removes fewer false pairs than a quality filter on the index reads. We therefore used a fixed minimum quality score filter of 15 on the index reads for all subsequent analyses.

Since we had stored the raw image data from the sequencing experiments, we were able to explore whether the occurrence of false index pairs changes if different versions of Illumina’s image analysis software (RTA/OLB 1.6, 1.8 and 1.9) are used. We found that the newer image analysis software identified a larger number of clusters, but also increased the fraction of falsely paired indexes (Supplementary Table S3). We also compared the performance of Illumina’s base caller Bustard to the base

caller Ibis (12), both of which directly operate on cluster intensity files. For both base callers, the fraction of false pairs increases if a larger number of correct pairs is identified, indicating an improved performance of the newer image analysis software in extracting signals from low quality clusters.

Disentangling the factors causing false index pairs

Oligonucleotide synthesis, amplification and sequencing errors are expected to create false index pairs at comparatively low frequencies. It can therefore be assumed that the vast majority of false index pairs remaining after quality filtering (0.060% in no-CAP, 0.138% in SP-CAP and 0.428% in MP-CAP) are caused by (i) remaining sequences from mixed clusters, (ii) cross-contamination of oligonucleotides or indexed libraries, or (iii) chimera formation during bulk amplification of libraries from different samples (MP-CAP only). These factors are expected to generate different patterns of false pairings, which may be used to further disentangle the underlying causes. The conversion of indexes due to mixed clusters and jumping PCR can be assumed to occur uniformly across all index pairs. In contrast, cross-contamination of indexes is expected to be a sporadic process and its effect size may be assessed from unusually frequent false pairs.

We therefore counted the occurrence of all possible index pairs in the three experiments (Figure 3). For identifying putative cross-contamination, we may identify false pairs with an overrepresentation of counts directly from the figure. Figure 3 clearly shows individual pairs, e.g. 11/103 and 11/105 in no-CAP, 97/3 and 10/105 SP-CAP, and 97/3 in MP-CAP, which are overrepresented compared to background. In addition, overrepresentation of complete rows/columns is also observed, e.g. the reverse index 1 in no-CAP and the forward index 106 in SP-CAP. Quantitatively, we checked for an overrepresentation of

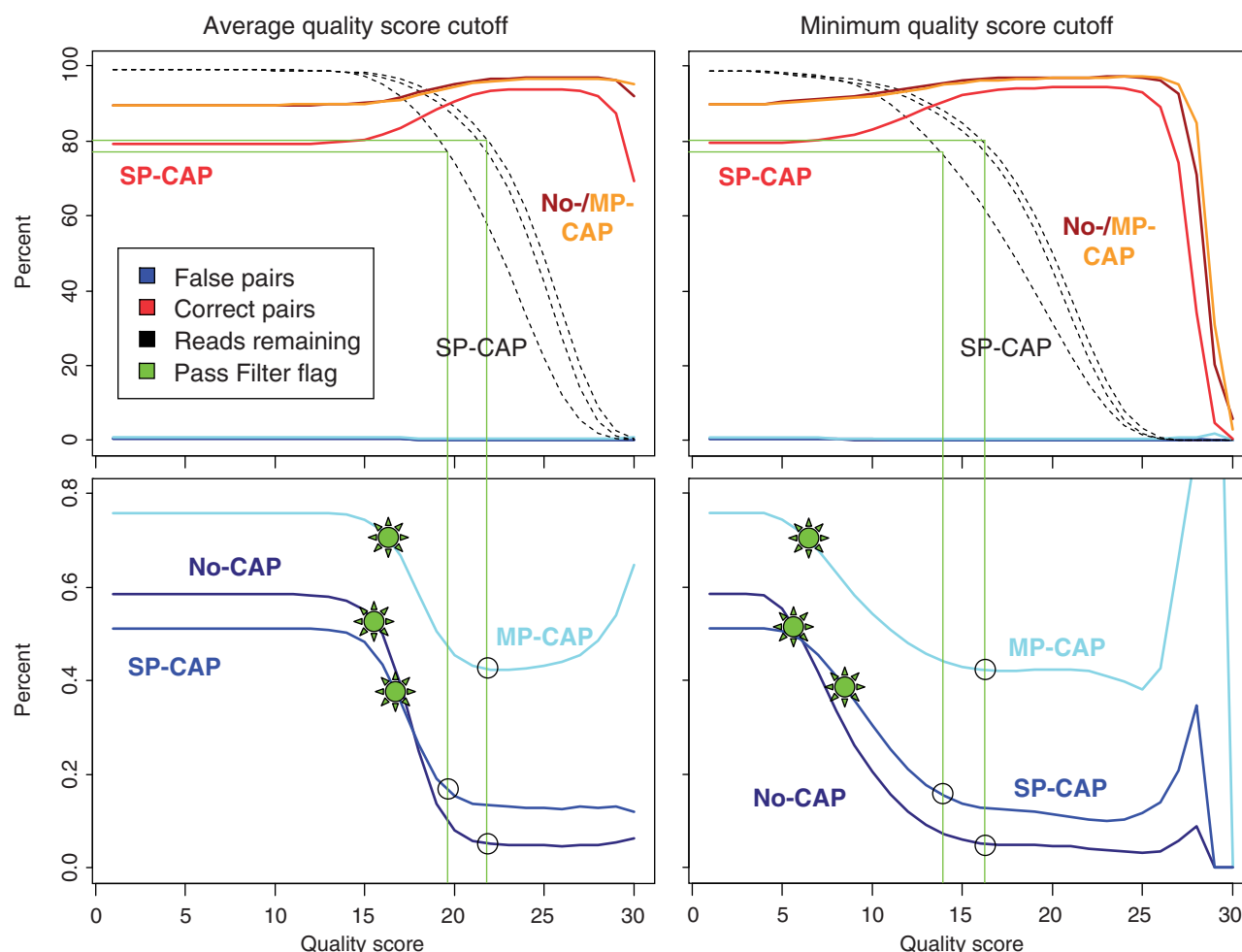


Figure 2. Changes in the fraction of false (blue) and correct (red) index pairs when applying two different types of base quality filters on the index reads (minimum accepted quality score and average quality score). The fraction remaining after PF is indicated by a green 'sun' symbol. Black circles denote the fraction of reads remaining when considering quality score cutoffs that remove just a little bit less raw data than the Pass Filter flag (green lines, ~20% of the data). Both filter criteria remove considerably more false pairs. While no-CAP, SP-CAP and MP-CAP show similar trajectories, the fraction of false pairs is always considerably higher for the MP-CAP experiment, in which samples have been enriched and amplified in a multiplex setup. Quality score cutoffs for the SP-CAP experiment are lower than for the other two experiments due to the 40% higher cluster density of this experiment.

false index combinations compared to a background value calculated for each experiment (see Supplementary Methods section). When estimating cross-contamination from index pairs that are observed five times more frequently than the background, thus considering only the higher frequency false index pairs, we estimate that 0.042% (no-CAP), 0.104% (SP-CAP) and 0.038% (MP-CAP) of the false pairs are due to cross-contamination. The comparatively high contamination estimate for the SP-CAP experiment is also supported by another—albeit less powerful—analysis, which can be performed by counting the number of sequences derived from unused indexes (24) (see Supplementary Methods section).

Subtracting the number of false pairs derived from cross-contamination from the total number of false pairs provides an estimate for the fraction of false pairs that are caused either by remaining mixed clusters or jumping PCR. Interestingly, these numbers are low for no-CAP

(0.018%) and SP-CAP (0.034%) and more than ten times higher for MP-CAP (0.390%). The low numbers in the first two experiments can be attributed to mixed cluster that could not be eliminated by quality filtering. The third experiment differs from the others in that libraries were amplified in bulk after target capture. We therefore conclude that jumping PCR generated ~0.36% chimeric reads with false index pairs in MP-CAP. If recombination happens predominantly along the adapter sequences, which are the regions of the sequencing library with the highest sequence similarity, half of the chimeric reads (0.18%) would be assigned to a false sample if only a single index read was used.

Quantifying false-assignment rates in single-indexed experiments

Assuming that both index reads are equally informative about sample origin, the false-assignment rate in experiment no-CAP would be 0.29% if only the first index was

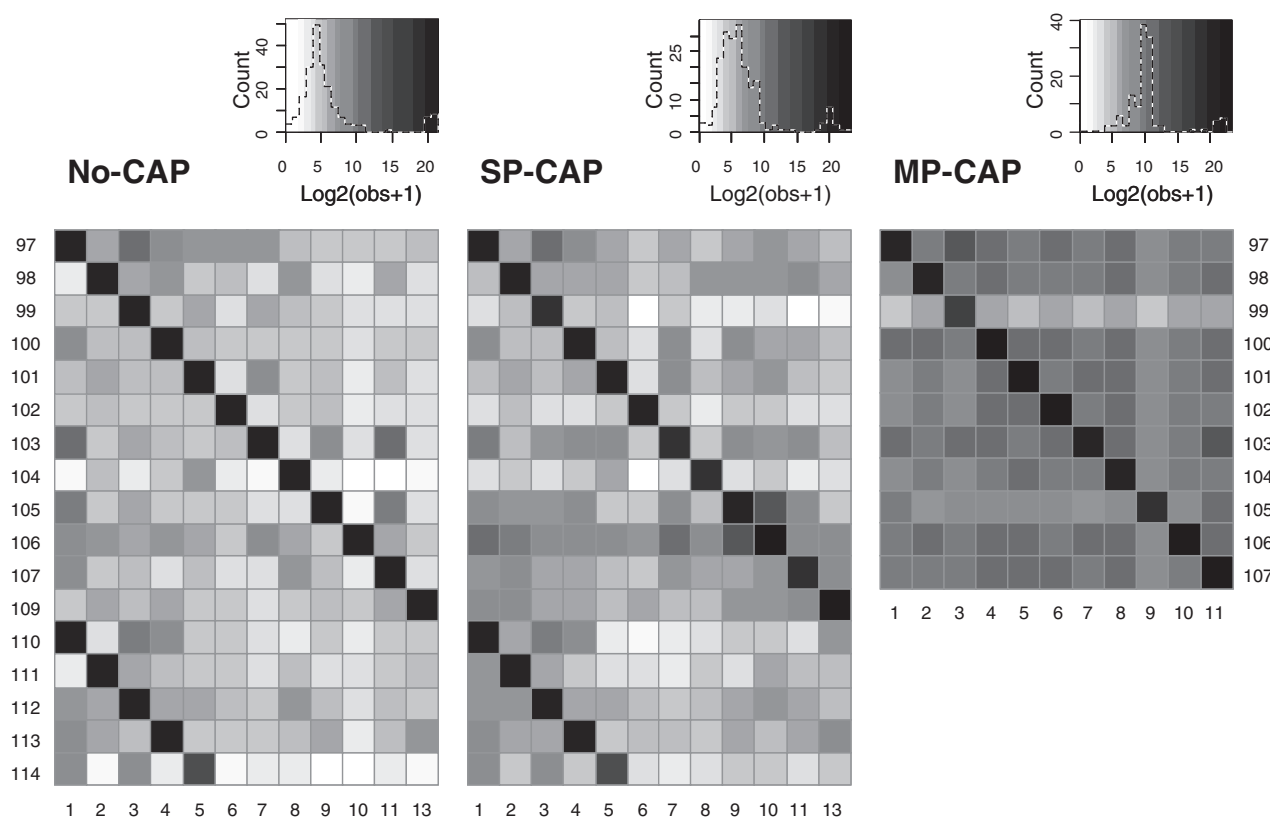


Figure 3. Heat map with the counts observed for all index combinations in the experiments no-CAP, SP-CAP and MP-CAP after applying a minimum quality score filter of 15 to the index reads. Only indexes that were actually used in the experiments are plotted; forward indexes on the horizontal and reverse indexes on the vertical axis. Color frequencies are provided for each of the experiments in the top right graphs.

used to sort the sequences (this is the fraction of sequences with falsely paired indexes divided by two). This number can almost exclusively be attributed to sequences from mixed clusters, demonstrating that false-assignments of sequences to samples occur at high frequency in single-indexed multiplex sequencing experiments, particularly if no quality filter is imposed on the index read. Although double-indexing is the most powerful approach for identifying and excluding sequences from mixed clusters, false-assignment rates may also be inferred from existing single-indexed data, if samples with a large evolutionary distance are sequenced together. In this case, sample identification based on the index read can be compared to sample identification based on alignments to the respective reference genomes.

In our three experiments—as in all our recent sequencing runs—we spiked a ϕ X174 control library into each lane of the flow cell (yielding ~1% of total reads). Sequences from this library are usually used to train the Ibis base caller (25) and to obtain measures of run quality for comparing different experiments. Unfortunately, ϕ X174 and Neandertal library sequences are not sufficiently distinct for this analysis, since bacteriophage sequences may also derive from microbial contaminants in the bone. We therefore analyzed raw data from seven human genomes we sequenced recently (26). These samples were indexed, but sequenced on separate lanes of one run. Based on the sequences identified as ϕ X174,

we determined false-assignment rates in the range of 0.09%–0.22% (see Supplementary Table S4 and Supplementary Methods section). However, these numbers are very likely underestimates, because the requirement of successful alignments implicitly acts as a quality filter. When applying the stringent minimum quality score filter of 15, false assignment rates reduce to 0.01%–0.03%. Finally, we also analyzed sequencing data from mRNA libraries, which were constructed using a very different library preparation protocol (Illumina's TruSeq RNA Sample Prep Kit, FC-122-100x), and found similar results. Between 0.14% and 0.17% of ϕ X174 reads erroneously occur with one of the sample indexes if no quality filter is applied to the index read (see Supplementary Table S5 and Supplementary Methods section). Thus, the high false-assignment rates we report do not represent artifacts of library preparation, but must be caused by the occurrence of mixed clusters on the flow cell.

DISCUSSION

Multiplex sequencing strategies have become indispensable for exploiting the capacities of high-throughput sequencing technologies in a cost- and time-efficient manner. However, little emphasis has been placed on directly assessing the level of confidence at which sequences are assigned to their source samples, probably

because these strategies are believed to be sufficiently accurate for most applications based on theoretical considerations. Sequences generated with our new double-indexing method reveal actual false-assignment of up to 0.3%, orders of magnitude higher than expected. The overwhelming majority of false assignments can be explained by mixed clusters, i.e. clusters originating from two different template molecules or clusters growing into each other. False assignment occurs if the dominance of signals changes in different reads or if different signals are tracked. Although we do not understand the exact underlying processes, we independently verified the existence of this effect in data from single-indexed multiplex sequencing experiments. Thus, it neither represents an artifact of the double-indexing method nor the library preparation protocols used.

In many cases, false-assignment rates in the order of 3 in 1000 sequences can be highly problematic. For example, rare somatic mutations are used as biomarkers for cancer (27–29) or for studying mitochondrial heteroplasmy (30,31). Gene expression studies may be confounded by the bleeding-over of sequences from one sample to another. High false-assignment rates may even be problematic for accurate genotyping in studies using targeted re-sequencing. If, for instance, enrichment success in hybridization capture varies among samples, sequences bleeding over from a highly enriched sample (e.g. a positive control) may eventually constitute several percent of the target sequences in a weakly enriched sample. Apart from mixed clusters, we identified two other major sources of error that lead to false assignments of sequences to samples. The first is sporadic cross-contamination of oligonucleotides carrying different indexes, which may be introduced during synthesis or subsequent handling step. Despite being cautious, we were not able to completely avoid this type of contamination in our experiments. The second, PCR jumping, occurs only in experiments where sequencing libraries from multiple samples are amplified in bulk, leading to a significant fraction of chimeric molecules (~0.4% in our experiment). Relative to these errors, amplification and sequencing errors, which were often focused on in previous studies, occur at negligible levels.

We developed two strategies to improve the accuracy of multiplex genotyping on the Illumina platform. The first and most powerful strategy is using our double-indexing method. Here, sample identification is performed twice for each template molecule, enabling an exponential decrease of the false-assignment rates. For example, by identifying and removing false index pairs in experiment MP-CAP, the false-assignment rate drops from ~2 in 1000 to less than 1 in 100 000. Thus, using double-indexing, accurate genotyping becomes possible even if libraries from different samples are amplified in bulk or if cross-contamination is present among indexed oligonucleotides. Moreover, double-indexing greatly reduces the costs of highly multiplexed sequencing. If only 50 indexed oligonucleotides are synthesized for each of the two adapters, 2500 index combinations are theoretically available. Since this level of multiplexing is hardly ever required, the majority of index combinations will remain unused.

This allows for determining false-assignment rates with nearly the same precision as if each index is only used once. Although double-indexing is recommended for all applications requiring bulk amplification of indexed libraries (e.g. multiplex target capture) or extraordinary levels of accuracy, we suggest a second—not mutually exclusive—strategy for reducing false-assignment rates also in single-indexed experiments, which is applying a quality filter on the index read. However, with this strategy it remains impossible to estimate false-assignment rates in single-indexed experiments, unless samples with a large evolutionary distance are sequenced together. Spiking in a ϕ X174 control library will often be suitable for this purpose.

Alternative indexing strategies for the Illumina platform have been developed where an index is attached immediately adjacent to the template molecule. Using these strategies, index and template are sequenced simultaneously in the forward read. In consequence, the confounding effect of mixed clusters can be expected to be much smaller. However, as with all single-indexed approaches, other sources of sample misidentification cannot be prevented, most notably oligonucleotide cross-contamination and jumping PCR. We conclude that incorporating indexes into both ends of library molecules is a very powerful approach for improving the accuracy of multiplex sequencing. This approach can in principle also be extended to other high-throughput sequencing platforms to reduce the errors common to multiplex sequencing in general and uncover problems inherent to the specific technology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Tables 1–5, Supplementary Figures 1–2, Supplementary Methods, Supplementary Reference (32).

ACKNOWLEDGEMENTS

The authors thank the members of the Department of Evolutionary Anthropology, in particular Janet Kelso and Svante Pääbo, for providing interesting discussions and useful insights.

FUNDING

Max Planck Society. Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

3. 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. *Nature*, **456**, 53–59.
4. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
5. Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B., Pham, T.T., Otto, G.A., Foquet, M. and Turner, S.W. (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl Acad. Sci. USA*, **105**, 1176–1181.
6. Meyer, M., Stenzel, U., Myles, S., Prufer, K. and Hofreiter, M. (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res.*, **35**, e97.
7. Craig, D.W., Pearson, J.V., Szlinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A. *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods*, **5**, 887–893.
8. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.
9. Meyer, M. and Kircher, M. (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc.*, **2010**, pdb prot5448.
10. Illumina Inc. (2008). 770-2008-011 ed http://www.illumina.com/Documents/products/datasheets/datasheet_sequencing_multiplex.pdf (11 September 2011, date last accessed).
11. Erlich, Y., Mitra, P.P., delaBastide, M., McCombie, W.R. and Hannon, G.J. (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods*, **5**, 679–682.
12. Kircher, M., Stenzel, U. and Kelso, J. (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.*, **10**, R83.
13. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., 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**, 153–158.
14. Green, R.E., Briggs, A.W., Krause, J., Prufer, K., Burbano, H.A., Siebauer, M., Lachmann, M. and Paabo, S. (2009) The Neandertal genome and ancient DNA authenticity. *EMBO J.*, **28**, 2494–2502.
15. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.
16. Stiller, M., Knapp, M., Stenzel, U., Hofreiter, M. and Meyer, M. (2009) Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Res.*, **19**, 1843–1848.
17. Meyerhans, A., Vartanian, J.P. and Wain-Hobson, S. (1990) DNA recombination during PCR. *Nucleic Acids Res.*, **18**, 1687–1691.
18. Paabo, S., Irwin, D.M. and Wilson, A.C. (1990) DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.*, **265**, 4718–4721.
19. 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**, 2049–2057.
20. Lahr, D.J. and Katz, L.A. (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques*, **47**, 857–866.
21. Maricic, T., Whitten, M. and Paabo, S. (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*, **5**, e14004.
22. Rohland, N. and Hofreiter, M. (2007) Comparison and optimization of ancient DNA extraction. *Biotechniques*, **42**, 343–352.
23. Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M. and Paabo, S. (2009) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.*, doi:10.1093/nar/gkp1163.
24. Meyer, M., Stenzel, U. and Hofreiter, M. (2008) Parallel tagged sequencing on the 454 platform. *Nat. Protoc.*, **3**, 267–278.
25. Kircher, M. and Kelso, J. (2010) High-throughput DNA sequencing—concepts and limitations. *Bioessays*, **32**, 524–536.
26. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L. *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053–1060.
27. Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L. *et al.* (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, **467**, 1109–1113.
28. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
29. Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.
30. Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I. and Stoneking, M. (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.*, **87**, 237–249.
31. He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz, L.A. Jr, Kinzler, K.W., Vogelstein, B. and Papadopoulos, N. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, **464**, 610–614.
32. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform

Martin Kircher, Susanna Sawyer, and Matthias Meyer

CONTENTS

SUPPLEMENTARY TABLES AND FIGURES

Supplementary Table 1. Primer sequences	2
Supplementary Table 2. Index combination used for each library	3
Supplementary Table 3: Fraction of false pairs observed when using different image analysis and base calling software	3
Supplementary Table 4: Quantifying false sample assignments in human sequencing data from single-indexed libraries	4
Supplementary Table 5: Quantifying false sample assignments in human and mouse transcriptome sequencing data from single-indexed libraries	4
Supplementary Figure 1: Raw intensity values of false index pairs	5
Supplementary Figure 2: Changes in the fraction of false index pairs	6

SUPPLEMENTARY METHODS 7

1. Quantifying false index pairs expected based on sequencing error alone	7
2. Analyzing intensity values from a random set of clusters with false index pairs	8
3. Reading each index twice from short-insert molecules	11
4. Quantifying jumping PCR and other effects	11
5. Estimating false-assignment rates based on the occurrence of unused indexes	16
5. Identification of false sample assignments in single indexed data	17

SUPPLEMENTARY REFERENCES..... 17

Supplementary Table 1. Primer sequences

All primers were synthesized by Sigma-Aldrich (Steinheim, Germany). Indexing primers were purified using reverse phase cartridges (RPC). All other primers were purified by HPLC.

Primer ID / index	Sequence (5' -> 3')
<i>P7 indexing primers (first index)</i>	
Sol_iPCR-MPI-97 AATCTTC	CAAGCAGAAGACGGCATACGAGATgaagattGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-98 ACCAACG	CAAGCAGAAGACGGCATACGAGATcggttggtGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-99 AGATGGC	CAAGCAGAAGACGGCATACGAGATgccatctGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-100 CCAGGTT	CAAGCAGAAGACGGCATACGAGATaacctggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-101 CCGTTAG	CAAGCAGAAGACGGCATACGAGATctaacggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-102 CGCTCT	CAAGCAGAAGACGGCATACGAGATagaggcgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-103 CTTGCGG	CAAGCAGAAGACGGCATACGAGATccgcaagGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-104 GGCGGAG	CAAGCAGAAGACGGCATACGAGATtccgccGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-105 TGGACGT	CAAGCAGAAGACGGCATACGAGATacgtccaGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-106 AACCATG	CAAGCAGAAGACGGCATACGAGATcatggttGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-107 CAGGAAG	CAAGCAGAAGACGGCATACGAGATcttcctgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-108 CATACTT	CAAGCAGAAGACGGCATACGAGATaggtatgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-109 CCAATCC	CAAGCAGAAGACGGCATACGAGATggattggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-110 CCGCGGT	CAAGCAGAAGACGGCATACGAGATacgccggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-111 CGCATAG	CAAGCAGAAGACGGCATACGAGATctatgcgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-112 CGTAATC	CAAGCAGAAGACGGCATACGAGATgattacgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-113 CGTTGGT	CAAGCAGAAGACGGCATACGAGATaccaacgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-114 CTATACG	CAAGCAGAAGACGGCATACGAGATcgtatagGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-115 GACCTAC	CAAGCAGAAGACGGCATACGAGATgtaggtcGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-116 GATATTG	CAAGCAGAAGACGGCATACGAGATcaatatcGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-117 AAGACGC	CAAGCAGAAGACGGCATACGAGATgcgtcttGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-118 GCAGTAT	CAAGCAGAAGACGGCATACGAGATatactgcGTGACTGGAGTTCAGACGTGT
Sol_iPCR- φX TTGCCGC	CAAGCAGAAGACGGCATACGAGATgcggaagGTGACTGGAGTTCAGACGTGT
<i>P5 indexing primers (second index)</i>	
P5_iPCR-LP-1 TCGCAGG	AATGATACGGCGACCACCGAGATCTACACcctgcgaACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-2 CTCTGCA	AATGATACGGCGACCACCGAGATCTACActgcagagACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-3 CCTAGGT	AATGATACGGCGACCACCGAGATCTACACacctaggACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-4 GGATCAA	AATGATACGGCGACCACCGAGATCTACActtgatccACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-5 GCAAGAT	AATGATACGGCGACCACCGAGATCTACACatcttgACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-6 ATGGAGA	AATGATACGGCGACCACCGAGATCTACActctccatACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-7 CTCGATG	AATGATACGGCGACCACCGAGATCTACACcatcgagACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-8 GCTCGAA	AATGATACGGCGACCACCGAGATCTACActtcgagcACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-9 ACCAACT	AATGATACGGCGACCACCGAGATCTACACagttggtACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-10 CCGGTAC	AATGATACGGCGACCACCGAGATCTACActgaccggACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-11 AACTCCG	AATGATACGGCGACCACCGAGATCTACACcggagtACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-12 TTGAAGT	AATGATACGGCGACCACCGAGATCTACACacttcaaACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-13 ACTATCA	AATGATACGGCGACCACCGAGATCTACActgatagtACACTCTTTCCCTACACGACGCTCTT
IS4 AGATCTC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
<i>amplification primers for indexed libraries</i>	
IS5	AATGATACGGCGACCACCGA
IS6	CAAGCAGAAGACGGCATACGA
<i>sequencing primer</i>	
P5 index sequencing	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Supplementary Table 2. Index combination used for each library

Experiment MP-CAP			Experiments no-CAP and SP-CAP					
Sample	Index 1	Index 2	Sample	Index 1	Index 2	Sample	Index 1	Index 2
L1	AATCTTC (97)	TCGCAGG (1)	L1	AATCTTC (97)	TCGCAGG (1)	L10	AACCATG (106)	CCGGTAC (10)
L2	ACCAACG (98)	CTCTGCA (2)	L2	ACCAACG (98)	CTCTGCA (2)	L11	CCAATCC (109)	ACTATCA (13)
L3	AGATGGC (99)	CCTAGGT (3)	L3	AGATGGC (99)	CCTAGGT (3)	L12	CCGGCGT (110)	TCGCAGG (1)
L4	CCAGGTT (100)	GGATCAA (4)	L4	CCAGGTT (100)	GGATCAA (4)	L13	CGCATAG (111)	CTCTGCA (2)
L5	CCGTTAG (101)	GCAAGAT (5)	L5	CCGTTAG (101)	GCAAGAT (5)	L14	CGTAATC (112)	CCTAGGT (3)
L6	CGCCTCT (102)	ATGGAGA (6)	L6	CGCCTCT (102)	ATGGAGA (6)	L15	CGTTGGT (113)	GGATCAA (4)
L7	CTTGCGG (103)	CTCGATG (7)	L7	CTTGCGG (103)	CTCGATG (7)	L16	CAGGAAG (107)	AACTCCG (11)
L8	GGCGGAG (104)	GCTCGAA (8)	L8	GGCGGAG (104)	GCTCGAA (8)	L17	CTATACG (114)	GCAAGAT (5)
L9	TGGACGT (105)	ACCAACT (9)	L9	TGGACGT (105)	ACCAACT (9)	φX	TTGCCGC (control)	AGATCTC (IS4)
L10	AACCATG (106)	CCGGTAC (10)						
L16	CAGGAAG (107)	AACTCCG (11)						
φX	TTGCCGC (control)	AGATCTC (IS4)						

Supplementary Table 3: Fraction of false pairs observed when using different image analysis and base calling software

The sequencing run was performed close to the release date of a new image analysis version and images were transferred off the instrument. Thus, image analysis of this sequencing run was done once with the Illumina RTA software version 1.6 (on the instrument) and once with Illumina OLB version 1.8. Later, the analysis was repeated using OLB 1.9. Results reported in the manuscript are all based on OLB 1.8, which identified between 18%-25% more clusters for the different lanes than the original instrument software run with RTA1.6/SCS2.6. Both new image analysis software versions increased the fraction of perfect index pairings in these lanes by 1-5% (using the same base caller). This indicates a lower sequencing error due to the improved identification and tracking of cluster positions in the images of the flow cell. However, when comparing the fraction of false index pairs, increased values are observed for the two new versions.

RTA / OLB	Raw clusters	Bustard				Ibis			
		Correct pairs		False pairs		Correct pairs		False pairs	
no-Cap									
1.6	29133246	24116308	83%	97315	0.40%	25091043	86%	126518	0.50%
1.8	34241955	26765784	78%	120236	0.45%	29950674	87%	175266	0.58%
1.9	34241993	29962477	88%	170564	0.57%	29873025	87%	177594	0.59%
SP-Cap									
1.6	38937256	27155457	70%	125147	0.46%	29042714	75%	143824	0.49%
1.8	48546372	34564560	71%	159982	0.46%	37690136	78%	192995	0.51%
1.9	48546297	37889964	78%	193461	0.51%	37426552	77%	202081	0.54%
MP-Cap									
1.6	29245166	24246039	83%	154789	0.63%	25187487	86%	179188	0.71%
1.8	34684183	28316453	82%	192938	0.68%	30261682	87%	230005	0.75%
1.9	34684220	30525097	88%	233016	0.76%	30159122	87%	230668	0.76%

Supplementary Table 4: Quantifying false sample assignments in human sequencing data from single-indexed libraries

All forward reads were aligned to the ϕ X174 genome using BWA. False sample assignments (FSA) are defined as reads showing a sample index but an alignment to the phage genome. QF columns show the changes after applying a minimum quality score filter of 15 to the index read.

HGDP ID	Raw sequences	Algn ϕ X index	Algn Sample index	FSA rate	Algn ϕ X index (QF)	Algn Sample index (QF)	FSA rate (QF)	Kept ϕ X index (QF)	Kept Sample index (QF)
HGDP00456	30562322	755869	1471	0.194%	743437	231	0.031%	98.36%	15.70%
HGDP00998	31188058	483305	409	0.085%	473923	56	0.012%	98.06%	13.69%
HGDP00665	34994524	781560	803	0.103%	768559	58	0.008%	98.34%	7.22%
HGDP00491	37133303	866342	943	0.109%	852285	49	0.006%	98.38%	5.20%
HGDP00711	39665263	758411	876	0.115%	738413	77	0.010%	97.36%	8.79%
HGDP01224	35608002	968626	1238	0.128%	956959	85	0.009%	98.80%	6.87%
HGDP00551	36576315	720162	1583	0.219%	691189	144	0.021%	95.98%	9.10%

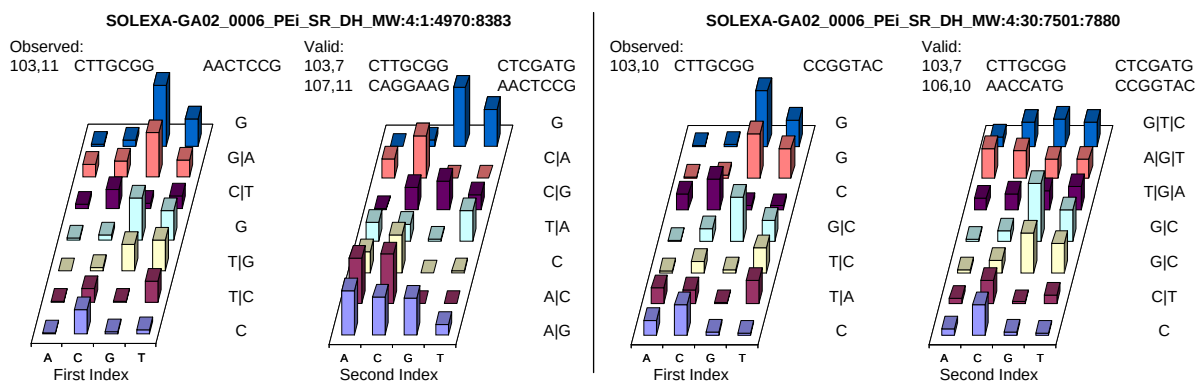
Supplementary Table 5: Quantifying false sample assignments in human and mouse transcriptome sequencing data from single-indexed libraries, which were prepared using Illumina's TruSeq RNA Sample Prep Kit

All single reads were aligned to the ϕ X174 genome using BWA. False sample assignments are defined as reads showing one of the 12 sample indexes but an alignment to the phage genome. QF columns show the changes after applying a minimum quality score filter of 15 to the index read.

Lane	Raw reads	Algn PhiX index	Ave. algn sample index	False sample assignments	Algn PhiX index QF	Ave. algn sample index QF	False sample assignments QF
1	45990409	563645	906	0.160%	551229	304	0.055%
2	42400537	611873	872	0.142%	600333	307	0.051%
3	44098398	482999	736	0.152%	474908	270	0.057%
4	43447369	597251	882	0.147%	586811	315	0.054%
5	50775674	412854	691	0.167%	404888	242	0.060%

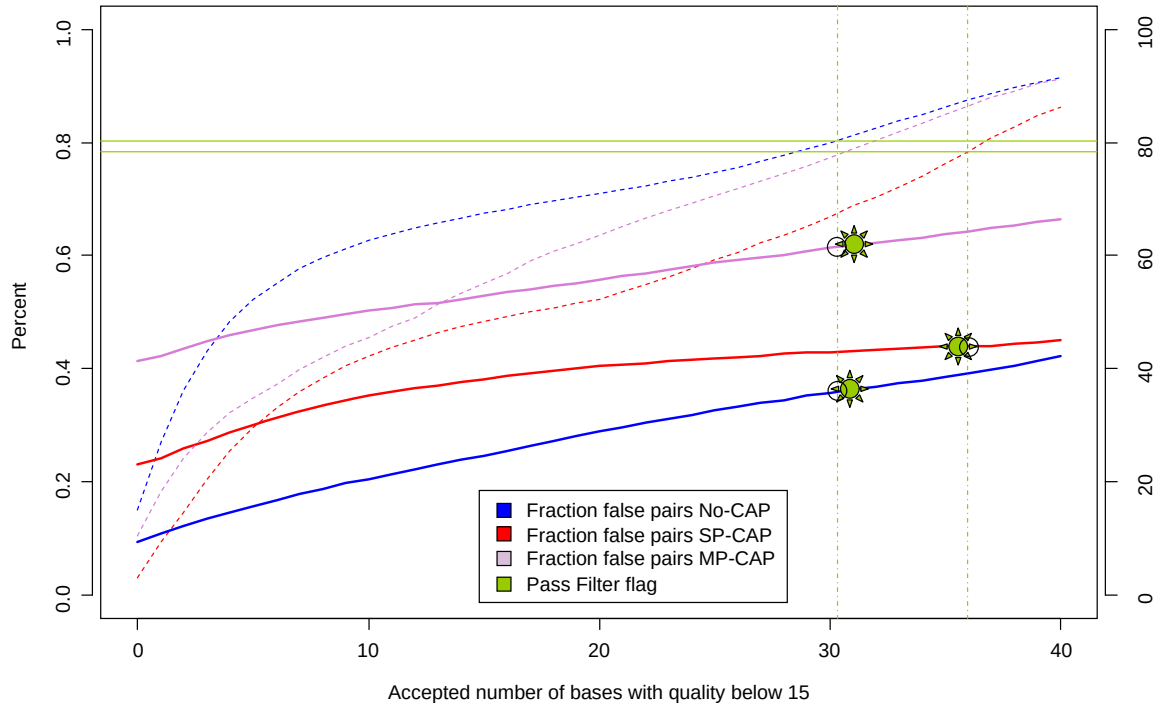
Supplementary Figure 1: Raw intensity values of false index pairs

We extracted the raw intensities from single clusters that were identified with a false index pair in each of the tiles 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 in the no-CAP experiment. The intensity values for both index reads of these twelve clusters are available in the Supplementary Text. All of them show intensity distributions indicative of non-pure clusters. Here, we show a visualization of the intensity values for tile 1 and 30. Illumina uses four fluorescent dyes to distinguish the four nucleotides A, C, G and T. Of these, two pairs (A/C and G/T) are excited using the same laser and are similar in their emission spectra. They are only partially separated using optical filters. A fluorophores are also measured in the C channel and G fluorophores are also measured in the T channel. Thus, an A can be identified by strong signals in both the A and the C channel, while a C will only show a strong signal in the C channel. The same applies for G/T. Hence, a strong signal in A/C excludes another signal in G/T channels for pure clusters. Hence, a strong signal in A/C excludes another signal in G/T channels for pure clusters.



Supplementary Figure 2: Changes in the fraction of false index pairs when accepting template reads with an increasing number of bases with quality scores smaller or equal to 15 (no-CAP blue, SP-CAP red, MP-CAP blue).

When considering a cutoff removing a little bit less raw data (dashed lines, right axis) than the Pass Filter flag (~20%; green lines), the applied filter removes slightly more false pairs in no-CAP and MP-CAP than the Pass Filter (PF) flag (sun symbols).



SUPPLEMENTARY METHODS

1. Quantifying false index pairs expected based on sequencing error alone

To obtain an estimate of false pairs from random errors, direct application of the binomial distribution provides an overestimate as usually only a very small proportion of the erroneous variants will match another index. In the presented experiments, at most 18 first indexes and 13 second indexes were used. We also required perfect matches to the index sequences and thus only a specific set of errors will generate a certain other valid index sequence. As a result only 17 and 12 of the 16,383 (4^7-1) erroneous variants, respectively, will contribute false index readouts. We can correct the estimate from the binomial distribution for this effect:

$$\sum_{x=1}^7 \frac{d_x}{\binom{7}{x} \cdot 3^x} \cdot \binom{7}{x} p^x (1-p)^{7-x}$$

The number of erroneous variants for a specific number of errors d_x can be inferred from the edit distance matrix of the 18 first index sequences:

```
[-, 6, 5, 6, 6, 5, 6, 7, 7, 3, 5, 7, 6, 4, 6, 6, 7, 5]
[6, -, 6, 6, 5, 5, 6, 5, 6, 3, 4, 6, 4, 5, 7, 5, 4, 7]
[5, 6, -, 5, 6, 6, 6, 5, 5, 6, 5, 6, 6, 5, 3, 7, 5, 5]
[6, 6, 5, -, 5, 5, 5, 5, 6, 6, 4, 3, 6, 5, 4, 5, 5, 7]
[6, 5, 6, 5, -, 5, 5, 5, 6, 6, 4, 4, 3, 6, 5, 3, 4, 6]
[5, 5, 6, 5, 5, -, 6, 5, 5, 5, 4, 5, 3, 5, 4, 6, 5, 6]
[6, 6, 6, 5, 5, 6, -, 5, 5, 6, 6, 3, 5, 5, 4, 4, 4, 4]
[7, 5, 5, 5, 5, 5, 5, -, 6, 5, 7, 6, 3, 6, 5, 4, 6, 7]
[7, 6, 5, 6, 6, 5, 5, 6, -, 7, 6, 3, 5, 5, 4, 6, 7, 3]
[3, 3, 6, 6, 6, 5, 6, 5, 7, -, 7, 7, 5, 5, 7, 4, 5, 6]
[5, 4, 5, 4, 4, 4, 6, 7, 6, 7, -, 5, 4, 4, 6, 6, 4, 6]
[7, 6, 6, 3, 4, 5, 3, 6, 3, 7, 5, -, 6, 6, 4, 4, 6, 4]
[6, 4, 6, 6, 3, 3, 5, 3, 5, 5, 4, 6, -, 4, 5, 4, 5, 7]
[4, 5, 5, 5, 6, 5, 5, 6, 5, 5, 4, 6, 4, -, 4, 5, 5, 6]
[6, 7, 3, 4, 5, 4, 4, 5, 4, 7, 6, 4, 5, 4, -, 6, 5, 6]
[6, 5, 7, 5, 3, 6, 4, 4, 6, 4, 6, 4, 4, 5, 6, -, 4, 6]
[7, 4, 5, 5, 4, 5, 4, 6, 7, 5, 4, 6, 5, 5, 5, 4, -, 6]
[5, 7, 5, 7, 6, 6, 4, 7, 3, 6, 6, 4, 7, 6, 6, 6, 6, -]
```

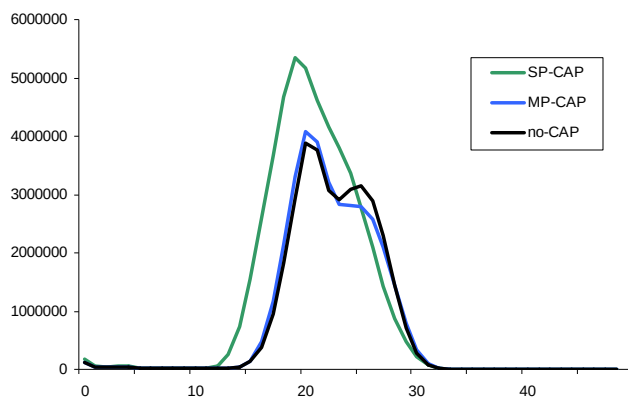
and the 13 indexes of the second index reads:

```
[-, 4, 5, 4, 6, 7, 7, 5, 4, 7, 7, 5, 6]
[4, -, 3, 6, 5, 6, 4, 5, 5, 4, 6, 7, 5]
[5, 3, -, 6, 5, 5, 4, 7, 5, 5, 4, 6, 6]
[4, 6, 6, -, 6, 6, 7, 6, 6, 6, 7, 3, 7]
[6, 5, 5, 6, -, 5, 6, 4, 5, 7, 6, 6, 4]
[7, 6, 5, 6, 5, -, 5, 5, 6, 5, 5, 6, 5]
[7, 4, 4, 7, 6, 5, -, 6, 5, 3, 4, 7, 5]
[5, 5, 7, 6, 4, 5, 6, -, 4, 7, 7, 7, 5]
[4, 5, 5, 6, 5, 6, 5, 4, -, 6, 5, 5, 7]
[7, 4, 5, 6, 7, 5, 3, 7, 6, -, 3, 4, 6]
[7, 6, 4, 7, 6, 5, 4, 7, 5, 3, -, 4, 5]
[5, 7, 6, 3, 6, 6, 7, 7, 5, 4, 4, -, 7]
[6, 5, 6, 7, 4, 5, 5, 5, 7, 6, 5, 7, -]
```

The matrix of 18 indexes indicates that on average 1.2 out of the 945 3-substitution variants, 3.1 out of the 2835 4-substitution variants, 5.6 out of the 5103 5-substitution variants, 5.3 out of the 5103 6-substitution variants and 1.8 out of 2187 7-substitution variants generate a valid other index of the forward index set. This corresponds to a d vector of (0, 0, 0, 1.2, 3.1, 5.6, 5.3, 1.8). For the second index set of 13 indexes, the d vector from the second matrix is (0, 0, 0, 0.6, 2.0, 3.7, 3.2, 2.5).

To apply the above binomial model, we also require an estimate of the average error rate for the three lanes. Such estimate can be obtained from the weighted average of error rates corresponding to the base quality scores ($10^{QS/10}$) of the raw reads. This way we obtain estimates of 1.022% raw sequencing error in no-CAP, 1.435% error in SP-CAP and 1.059% error in MP-CAP. The higher error rate in SP-CAP is due to the higher loading density of this lane (see Supplementary Figure S3). Considering the maximum error rate of 1.435% for p in the above equation results in $1.26E-07$ false index read outs for the first index set and $6.30E-08$ for the second index set.

Supplementary Figure S3: Distribution of base quality scores in the index reads



Adding the two rates for the forward and reverse set gives an upper estimate for false pairs of $1.89E-07$. This is an upper estimate as indexes are not independent in processing and both have to present a known index for the read to be considered. Direct application of the binomial model for a 1.435% error rate and without correcting for the number of valid erroneous index readouts would yield a rate of $9.90E-05$ for single indexes and $1.98E-04$ for index pairs.

2. Analyzing intensity values from a random set of clusters with false index pairs

We extracted the raw intensities from a single cluster that was identified with a false index pair in each of the tiles 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 in the no-CAP experiment. The intensity values for both index reads of these twelve clusters are available in the table below. Illumina's software reports intensities as intensity minus surrounding noise. Thus, if a large signal is observed around the cluster, a negative intensity can be reported. Illumina uses four fluorescent dyes to distinguish the four nucleotides A, C, G and T. Of these, two pairs (A/C and G/T) are excited using the same laser and are similar in their emission spectra. They are only partially separated using optical filters. A fluorophores are also measured in the C channel and G fluorophores are

also measured in the T channel. Thus, an A can be identified by strong signals in both the A and the C channel, while a C will only show a strong signal in the C channel. The same applies for G/T. Hence, a strong signal in A/C excludes another signal in G/T channels for pure clusters. Even though more complex intensity distributions can also hint to non-pure clusters, we simply marked the cycles where intensities of at least 200 are observed in both the A/C and G/T channels. We found at least one such observation in each of the two twelve clusters.

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:1:4970:8383									
103,11 CTTGCGG,AACTCCG									
Index1	A	C	G	T	Index2	A	C	G	T
1	39	1138	63	176	1	1407	1205	1176	347
2	65	674	67	1005	2	1452	1574	15	-44
3	0	188	1266	1502	3	645	1205	-34	52
4	99	235	2025	1398	4	602	539	66	972
5	223	889	238	613	5	45	690	920	483
6	604	792	2154	827	6	617	1340	-70	-13
7	100	288	2920	1284	7	20	202	1879	1193

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:10:5818:20510									
104,9 GCGGAG,ACCAACT									
Index1	A	C	G	T	Index2	A	C	G	T
1	-152	-244	212	97	1	1335	1304	284	114
2	-38	-54	164	56	2	103	918	69	9
3	-120	120	-49	-124	3	266	1129	52	139
4	-85	-3	162	10	4	1173	1229	64	25
5	26	-26	186	-5	5	1257	1151	175	93
6	119	24	-75	-15	6	271	893	-35	-15
7	20	-42	117	58	7	70	192	-52	988

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:20:6586:18234									
103,9 CTTGCGG,ACCAACT									
Index1	A	C	G	T	Index2	A	C	G	T
1	117	793	45	424	1	709	653	155	-15
2	-41	183	894	1177	2	91	428	83	95
3	78	196	906	1375	3	92	470	-2	186
4	670	523	1827	942	4	576	630	20	180
5	192	1075	270	354	5	494	451	89	124
6	64	290	2429	1095	6	16	408	132	168
7	111	279	1849	982	7	203	233	388	603

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:30:7501:7880									
103,10 CTTGCGG,CCGGTAC									
Index1	A	C	G	T	Index2	A	C	G	T
1	703	1466	139	107	1	161	731	80	49
2	777	676	84	1121	2	138	570	46	221
3	73	546	115	1209	3	70	291	940	690
4	40	602	2108	994	4	37	238	1383	742
5	735	1480	222	175	5	268	386	450	561
6	82	143	2113	1402	6	719	654	469	451
7	89	119	2694	1264	7	238	586	656	590

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:40:7680:4448									
109,9 CCAATCC,ACCAACT									
Index1	A	C	G	T	Index2	A	C	G	T
1	678	1180	86	224	1	462	391	8	254
2	593	1245	341	204	2	303	666	48	-1
3	1163	1090	391	557	3	70	370	299	405
4	1514	1557	114	2	4	530	780	0	96
5	72	218	92	1370	5	721	647	52	369
6	196	838	392	585	6	196	387	605	398
7	299	1214	60	156	7	161	258	503	570

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:50:9834:15557									
109,9 CCAATCC,ACCAACT									
Index1	A	C	G	T	Index2	A	C	G	T
1	55	227	63	59	1	1851	1757	50	63
2	119	326	38	66	2	219	1211	29	29
3	305	392	76	58	3	238	1121	-11	227
4	289	437	48	10	4	1618	1595	38	64
5	83	99	34	110	5	1358	1287	61	239
6	99	344	-34	44	6	168	1087	7	32
7	110	357	33	49	7	305	284	127	1217

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:60:9566:5016									
Index1	A	C	G	T	Index2	A	C	G	T
1	393	1273	73	430	1	563	644	413	189
2	67	306	686	1101	2	68	433	147	130
3	459	445	782	1279	3	268	567	62	157
4	413	451	2282	1239	4	476	648	141	162
5	198	1218	824	425	5	450	724	329	219
6	99	182	2793	1349	6	277	615	331	199
7	100	231	1856	1217	7	268	294	250	665

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:70:7191:7553									
Index1	A	C	G	T	Index2	A	C	G	T
1	88	1148	71	66	1	664	568	61	518
2	693	683	-173	723	2	788	924	69	96
3	-111	-52	1081	1315	3	152	582	687	305
4	26	14	2614	1080	4	327	356	252	731
5	674	1206	55	54	5	141	526	-33	418
6	588	566	1482	663	6	160	528	647	395
7	75	16	2407	987	7	12	35	2055	924

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:80:6406:10544									
Index1	A	C	G	T	Index2	A	C	G	T
1	78	299	98	80	1	2111	1864	92	95
2	388	414	84	27	2	557	1408	25	-9
3	51	54	668	318	3	284	1391	16	112
4	41	57	594	247	4	1428	1442	74	280
5	268	298	92	102	5	1612	1545	42	86
6	353	364	129	94	6	261	1147	28	50
7	72	110	610	257	7	51	58	575	1572

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:90:7464:1488									
Index1	A	C	G	T	Index2	A	C	G	T
1	114	431	49	54	1	1594	1421	-7	20
2	580	494	-8	80	2	463	1050	61	48
3	123	98	699	226	3	190	947	-5	197
4	121	146	653	241	4	843	776	49	315
5	450	349	44	129	5	868	1042	4	58
6	425	447	72	27	6	160	818	16	47
7	87	81	504	261	7	19	11	615	1119

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:100:5684:7314									
Index1	A	C	G	T	Index2	A	C	G	T
1	-1	-22	62	559	1	1567	1600	111	88
2	-113	31	835	436	2	175	1074	79	22
3	37	39	958	347	3	142	440	-20	999
4	562	590	24	24	4	1557	1524	33	-38
5	25	445	117	-9	5	517	519	86	1015
6	-66	-57	1187	531	6	258	1076	30	59
7	47	93	22	196	7	1130	1066	107	382

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:110:6123:5197									
Index1	A	C	G	T	Index2	A	C	G	T
1	-39	-7	866	392	1	1756	1500	575	341
2	55	87	718	309	2	267	1439	36	-14
3	94	217	68	48	3	37	31	104	1642
4	43	18	675	324	4	1487	1457	16	31
5	-83	37	624	318	5	77	86	734	1610
6	290	279	236	120	6	498	1315	16	67
7	98	120	564	203	7	1551	1502	107	103

SOLEXA-GA02_0006_Pei_SR_DH_MW:4:120:4821:17030									
Index1	A	C	G	T	Index2	A	C	G	T
1	277	361	40	43	1	1325	1181	216	90
2	394	418	-38	-21	2	308	944	-11	14
3	-22	123	301	96	3	286	893	414	285
4	110	276	-9	5	4	1019	967	367	383
5	323	269	64	61	5	874	748	479	263
6	207	190	39	224	6	365	707	344	81
7	39	12	635	264	7	44	180	476	886

3. Reading each index twice from short-insert molecules

We searched the sequence data of the three experiments for short-insert molecules where the complete and error-free adapter sequences with perfectly matching indexes were obtained in both the forward and reverse read. We identified a total of 3,574,203, 1,699,585 and 3,451,555 such clusters in no-CAP, SP-CAP and MP-CAP. For no-CAP, in 411 out of 3,574,203 ($11.499\text{E-}5$) of these observations, the indexes identified in the read out of the first index did not agree with its second read out. In 245 out of 3,574,203 ($6.855\text{E-}5$) the indexes read out from the second index did not agree. We obtained similar rates for SP-CAP (Index1: $209/1,699,585 = 12.297\text{E-}5$, Index2: $138/1,699,585 = 8.120\text{E-}5$) and MP-CAP (Index1: $305/3,451,555 = 8.837\text{E-}5$, Index2: $275/3,451,555 = 7.967\text{E-}5$). In cases where the actual index reads provided conflicting information on sample origin in the no-CAP experiment ($n=311$), the index read from the template read produced a valid index pair. Similar results were obtained for the two other experiments (SP-CAP 141 out of 149 observations, MP-CAP 99 out of 109).

4. Quantifying jumping PCR and other effects

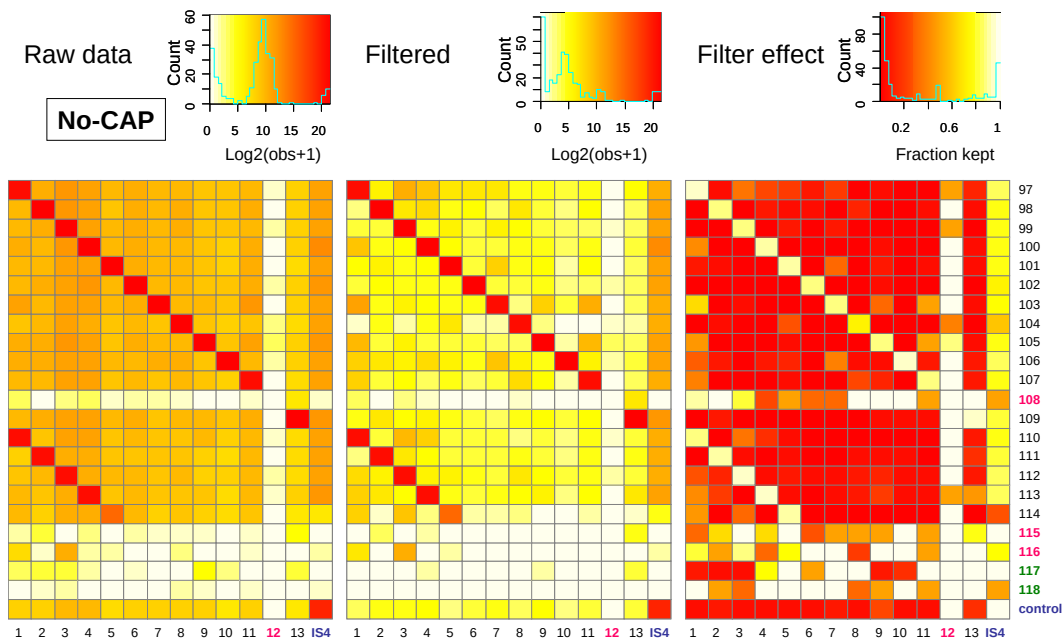
We checked the frequency of all putative index pairs, also including (i) index sequences that have not participated in the experiment at all, but were synthesized in the same batch of indexed oligonucleotides, and (ii) indexes that participated in some step of the experiment but did not participate in the final pooling (see Supplementary Figures 4-6 below). In all three experiments we see a clear increase of correct pairs compared to background when filtering the raw data based on the index reads, i.e. excluding all clusters where at least one base in any of the two index reads has a base quality score below 15. The quality filter affects false index pairs stronger than correct pairs.

When comparing row/column medians to the median of the medians for rows/columns (Supplementary Figures 4-6 below), in some rows and columns we see overrepresentation of false index pairs compared to background. Considering the median values assumes that always a minority of index pairs is affected by cross-contamination and that libraries are pooled in equimolar ratio. The row and column median of medians are 26.5/24, 61/60, and 1208.25/1050 for no-CAP, SP-CAP, and MP-CAP, respectively. IS4 represents the index sequence obtained from the P5-adapter of single-indexed libraries (i.e. using primer IS4 for the preparation of single index libraries instead of a second indexed oligonucleotide). This sequence is expected for the ϕX174 library, which was spiked into all sequencing lanes. IS4 shows a low reduction of counts when applying the quality filter to all three datasets and has much higher counts compared to the other rows. The IS4 medians are 1667, 2444 and 9055.5 for no-CAP, SP-CAP and MP-CAP respectively. This indicates contamination of the preparation chemicals (e.g. the PCR buffer) with the IS4 oligonucleotide from the Meyer and Kircher protocol (1). From no-CAP and SP-CAP, we estimate 0.11% and 0.25% contamination with IS4 correspondingly. Other examples of putative contamination include the second index 1 in no-CAP (5.8x higher than the median of medians) and the first index 106 in SP-CAP (10.6x higher than the median of medians). The figures clearly show individual pairs, e.g.

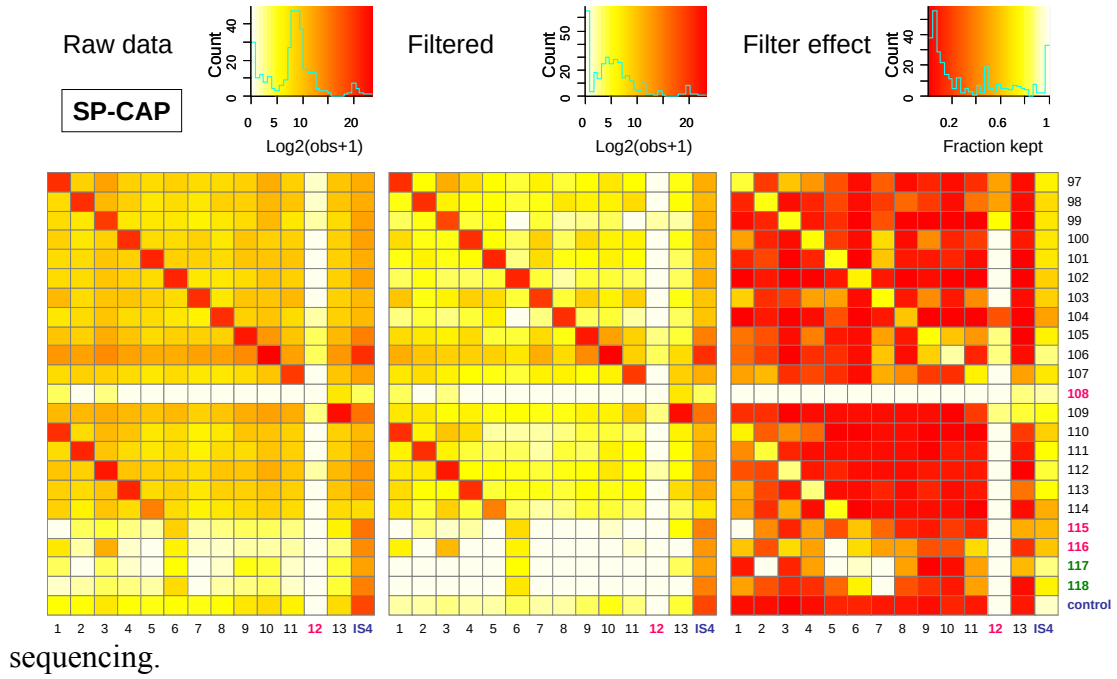
11/103 and 11/105 in no-CAP, 97/3 and 10/105 in SP-CAP, and 97/3 in MP-CAP that are overrepresented compared to background.

Indexes that were used in control reactions but not sequenced are seen 2-56 times more frequently in no-CAP than the ones not used in the experiment at all. This suggests handling contamination as one source of error. Assuming that higher frequency false index pairs (frequency five times higher than background) are due to contamination, we estimate that 0.04% of the 0.06% false pairs in no-CAP, 0.10% of the 0.14% false pairs in SP-CAP and 0.04% of the 0.43% false pairs in MP-CAP are due to cross-contamination of indexed oligonucleotides or libraries (Supplementary Tables 6.1-3 on the following pages).

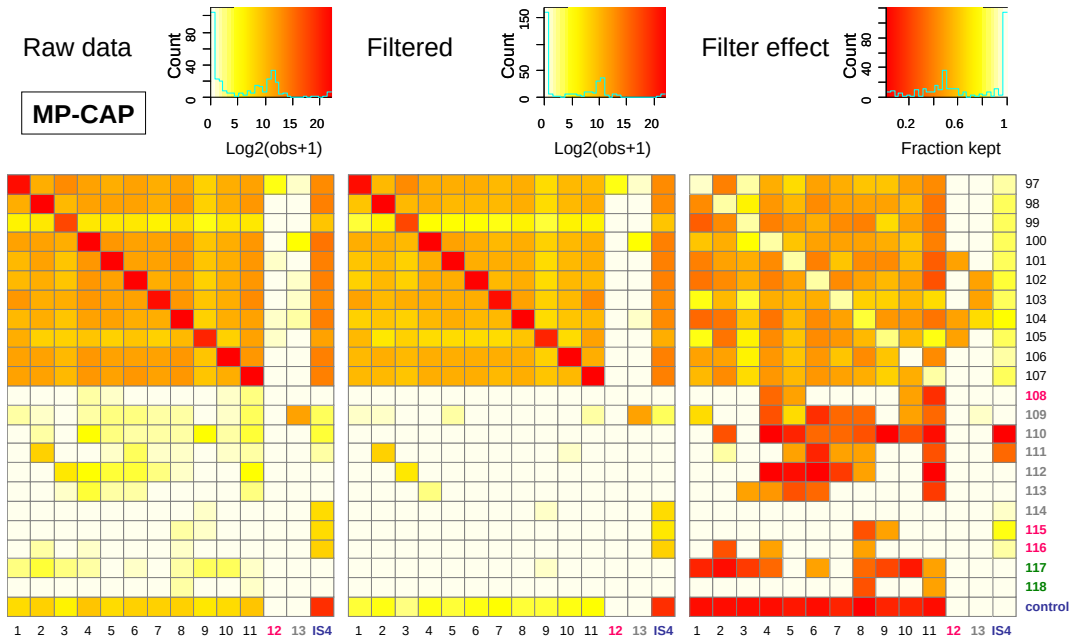
Supplementary Figure 4: Counts of all index pairs before and after applying a minimum base quality score cutoff of 15 to the index reads. Shown are the result for experiment no-CAP, where libraries were amplified independently and pooled just prior to sequencing. The first index is plotted on the vertical axis, the second on the horizontal axis. Indexes with green labels did not participate in the experiment at all. Pink indexes were used during library preparation, but not included in the library pool for sequencing. The control library is identified by the combination of *control* and *IS4* (blue).



Supplementary Figure 5: Results for perfect index pairings in experiment SP-CAP, where all libraries were enriched and amplified individually and just pooled prior to



Supplementary Figure 6: Results for perfect index pairings in experiment MP-CAP, where all libraries were pooled prior to enrichment and amplification.



Supplementary Table 6.1: Index cross-contamination in no-CAP quantified from higher frequency false index pairs (five times above the medians of the row and column medians). The upper part of the table gives the counts for all false index pairs, while in the lower table (next page) the higher frequency values have been replaced by the average of the medians of row/column medians.

Index	1	2	3	4	5	6	7	8	9	10	11	13	RowMd
97		60	1309	466	118	116	129	27	22	13	24	39	60
98	5		85	171	22	43	9	130	12	5	49	8	22
99	15	15		19	61	12	67	33	13	8	10	7	15
100	534	30	33		33	17	17	21	21	8	23	18	21
101	47	57	30	43		11	316	22	38	3	45	8	38
102	16	37	22	20	24		12	16	25	4	10	7	16
103	1705	19	79	40	21	28		6	262	12	1372	10	28
104	1	28	3	20	131	3	2		4	0	0	2	3
105	717	15	62	16	16	16	9			2	892	9	16
106	314	96	52	148	67	24	389	51	22		82	17	67
107	303	18	32	10	30	8	12	161	46	3		10	18
109	19	91	35	76	22	17	8	22	43	18	49		22
110		12	796	337	17	13	12	4	13	5	24	13	13
111	5		85	31	16	15	11	30	11	6	20	28	16
112	139	75		65	69	30	15	96	41	12	28	15	41
113	259	83	32		20	14	23	37	75	5	39	126	37
114	293	1	439	5		2	3	4	0	0	3	1	3
ColMedian	139	30	52	40	24	15.5	12	24.5	22	5	26	10	

Total count 16049
Total number of correct pairs 26901906
0.060%

Index	1	2	3	4	5	6	7	8	9	10	11	13
97		60	22.63	22.63	118	116	22.63	27	22	13	24	39
98	5		85	22.63	22	43	9	22.63	12	5	49	8
99	15	15		19	61	12	67	33	13	8	10	7
100	22.63	30	33		33	17	17	21	21	8	23	18
101	47	57	30	43		11	22.63	22	38	3	45	8
102	16	37	22	20	24		12	16	25	4	10	7
103	22.63	19	79	40	21	28		6	22.63	12	22.63	10
104	1	28	3	20	22.63	3	2		4	0	0	2
105	22.63	15	62	16	16	16	9	9		2	22.63	9
106	22.63	96	52	22.63	67	24	22.63	51	22		82	17
107	22.63	18	32	10	30	8	12	22.63	46	3		10
109	19	91	35	76	22	17	8	22	43	18	49	
110		12	22.63	22.63	17	13	12	4	13	5	24	13
111	5		85	31	16	15	11	30	11	6	20	28
112	22.63	75		65	69	30	15	96	41	12	28	15
113	22.63	83	32		20	14	23	37	75	5	39	22.63
114	22.63	1	22.63	5		2	3	4	0	0	3	1

Total count 4777
Cross-contamination corrected 0.018%

Supplementary Table 6.2: Quantification of index cross-contamination in SP-CAP from higher frequency false index pairs.

Index	1	2	3	4	5	6	7	8	9	10	11	13	RowMd
97		68	2536	277	64	22	76	14	65	150	69	42	68
98	34		87	58	48	18	39	132	126	104	251	59	59
99	12	26		15	25	0	23	4	5	10	0	2	10
100	230	25	25		44	10	429	11	312	78	82	37	44
101	28	78	29	61		7	362	29	55	105	47	13	47
102	9	28	12	8	21		15	4	15	21	8	8	12
103	828	41	119	410	249	16		17	409	95	524	30	119
104	6	19	7	21	88	0	7		9	23	3	10	9
105	359	160	122	348	24	21	269	30		5623	468	17	160
106	2223	780	490	646	474	148	3026	238	9666		583	353	583
107	154	328	84	74	33	16	101	86	65	120		427	86
109	194	193	49	77	35	61	32	27	103	103	456		77
110		80	693	259	4	2	4	6	20	21	10	99	20
111	119		91	15	10	10	4	14	11	79	44	43	15
112	171	128		65	62	34	14	37	74	172	58	47	62
113	247	59	57		15	29	26	18	60	38	48	305	48
114	208	17	509	16		9	3	7	17	24	12	13	16
ColMedian	171	68	87	65	35	16	29	17.5	62.5	87	53	39.5	

Total count 42761
Total number of correct pairs 30989801
0.138%

Index	1	2	3	4	5	6	7	8	9	10	11	13
97		68	52.9	277	64	22	76	14	65	150	69	42
98	34		87	58	48	18	39	132	126	104	251	59
99	12	26		15	25	0	23	4	5	10	0	2
100	230	25	25		44	10	52.9	11	52.9	78	82	37
101	28	78	29	61		7	52.9	29	55	105	47	13
102	9	28	12	8	21		15	4	15	21	8	8
103	52.9	41	119	52.9	249	16		17	52.9	95	52.9	30
104	6	19	7	21	88	0	7		9	23	3	10
105	52.9	160	122	52.9	24	21	269	30		52.9	52.9	17
106	52.9	52.9	52.9	52.9	52.9	148	52.9	238	52.9		52.9	52.9
107	154	52.9	84	74	33	16	101	86	65	120		52.9
109	194	193	49	77	35	61	32	27	103	103	52.9	
110		80	52.9	259	4	2	4	6	20	21	10	99
111	119		91	15	10	10	4	14	11	79	44	43
112	171	128		65	62	34	14	37	74	172	58	47
113	247	59	57		15	29	26	18	60	38	48	52.9
114	208	17	52.9	16		9	3	7	17	24	12	13

Total count 10622
Cross-contamination corrected 0.034%

Supplementary Table 6.3: Quantification of index cross-contamination in MP-CAP from higher frequency false index pairs.

Index	1	2	3	4	5	6	7	8	9	10	11	RowMd
97		727	6949	1328	1206	1325	819	1885	184	894	1196	1201
98	621		781	1625	1167	965	941	1826	192	1135	1377	1050
99	19	63		52	46	51	30	68	17	60	58	51.5
100	1283	1277	976		2269	1694	1055	1794	290	1092	1931	1280
101	471	1141	363	1761		1211	1707	1596	253	944	1562	1176
102	425	747	345	1480	1284		837	1374	224	786	1037	811.5
103	2933	996	1326	1989	1215	1385		2103	363	833	5659	1355.5
104	327	704	425	951	1344	842	730		212	697	1098	717
105	1012	156	262	219	246	244	174	447		233	2522	245
106	1087	1667	880	1578	1733	1211	1645	1902	285		1891	1611.5
107	1142	1068	652	1916	1467	1626	1054	1976	461	1013		1105
ColMedian	816.5	871.5	716.5	1529	1249.5	1211	889	1810	238.5	863.5	1469.5	

Total count 118717
Total number of correct pairs 27636356
0.428%

Index	1	2	3	4	5	6	7	8	9	10	11
97		727	997	1328	1206	1325	819	1885	184	894	1196
98	621		781	1625	1167	965	941	1826	192	1135	1377
99	19	63		52	46	51	30	68	17	60	58
100	1283	1277	976		2269	1694	1055	1794	290	1092	1931
101	471	1141	363	1761		1211	1707	1596	253	944	1562
102	425	747	345	1480	1284		837	1374	224	786	1037
103	2933	996	1326	1989	1215	1385		2103	363	833	997
104	327	704	425	951	1344	842	730		212	697	1098
105	1012	156	262	219	246	244	174	447		233	2522
106	1087	1667	880	1578	1733	1211	1645	1902	285		1891
107	1142	1068	652	1916	1467	1626	1054	1976	461	1013	
Total count											108103
Cross-contamination corrected											0.390%

5. Estimating false-assignment rates based on the occurrence of unused indexes

In standard multiplex sequencing experiments with single-indexed libraries, false-assignment rates can only be estimated by quantifying the occurrence of unused index sequences (2). These unused indexes are expected to appear if indexes are converted into each other due to errors in synthesis, amplification and sequencing or if there is cross-contamination among index PCR primers and indexed libraries. When we restrict our analysis to the forward index read, we determine false-assignment rates of 0.02% in no-CAP, 0.68% in SP-CAP and 0.004% in MP-CAP from five unused first indexes. We also performed the reciprocal analysis and analyzed perfect reverse index reads for which one index primer was not used. Here we obtained false-assignment rates of 0.001% for no-CAP, SP-CAP and MP-CAP, respectively. These rates can be combined to an estimate for false pairs by considering the average number of sequences observed for a first and a second index as well the expected number of sequences observed for a sample:

$$\frac{\overline{\#SeqIndex1} \cdot r_1 + \overline{\#SeqIndex2} \cdot r_2}{(\sum \#SeqIndex1 + \sum \#SeqIndex2) / (2 \cdot \#Samples)}$$

Using this formula, we obtain the joint estimates of:

	no-CAP	SP-CAP	MP-CAP
Unused forward indexes (r_1)	0.020%	0.681%	0.004%
Unused reverse index (r_2)	0.001%	0.001%	0.001%
Number of samples	17	17	11
Average per used first index	2125542	3381013	2888431
Average per used second index	2201444	2458359	2887183
Sum used first indexes	31364391	41792106	31772738
Sum used second indexes	31566651	40572152	31759009
Joint estimate unused indexes	0.024%	0.818%	0.005%

The false-assignment rates estimated from unused indexes is highest in SP-CAP. Since amplification and sequencing errors will only rarely convert one index sequence into another (see main text), these rates must almost exclusively reflect cross-contamination among indexed oligonucleotides or libraries.

5. Identification of false sample assignments in single indexed data

To quantify false index assignments in regular single index libraries, we used the 2x101+7 PE sequencing data from 7 present-day humans presented by Reich et al.(3). From the seven lanes, we separately aligned the forward and reverse read of all raw clusters to the ϕ X174 reference genome using BWA(4) and identified false index assignments as reads with a ϕ X alignment showing the designated sample index for the specific library. Supplementary Table 2 provides the results for the forward read with and without applying a minimum quality score filter of 15 to the index read out. In a second experiment of five lanes with mRNA libraries, generated using the Illumina TruSeq RNA Sample Prep Kit and sequenced on a single read run with 76+7 cycles, we obtain on average 0.14% to 0.17% ϕ X174 contamination for each sample index, which reduces to 0.05-0.06% after applying the index quality filter (Supplementary Table 3).

SUPPLEMENTARY REFERENCES

1. Meyer, M. and Kircher, M. (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc*, **2010**, pdb prot5448.
2. Meyer, M., Stenzel, U. and Hofreiter, M. (2008) Parallel tagged sequencing on the 454 platform. *Nat Protoc*, **3**, 267-278.
3. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L. *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053-1060.
4. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.