Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform

Martin Kircher, Susanna Sawyer, and Matthias Meyer

CONTENTS

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')
P7 indexing primers (first index)	
Sol_iPCR-MPI-97 AATCTTC	CAAGCAGAAGACGGCATACGAGATgaagattGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-98 ACCAACG	CAAGCAGAAGACGGCATACGAGATcgttggtGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-99 AGATGGC	CAAGCAGAAGACGGCATACGAGATgccatctGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-100 CCAGGTT	CAAGCAGAAGACGGCATACGAGATaacctggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-101 CCGTTAG	CAAGCAGAAGACGGCATACGAGATctaacggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-102 CGCCTCT	CAAGCAGAAGACGGCATACGAGATagaggcgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-103 CTTGCGG	CAAGCAGAAGACGGCATACGAGATccgcaagGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-104 GGCGGAG	CAAGCAGAAGACGGCATACGAGATctccgccGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-105 TGGACGT	CAAGCAGAAGACGGCATACGAGATacgtccaGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-106 AACCATG	CAAGCAGAAGACGGCATACGAGATcatggttGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-107 CAGGAAG	CAAGCAGAAGACGGCATACGAGATcttcctgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-108 CATACCT	CAAGCAGAAGACGGCATACGAGATaggtatgGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-109 CCAATCC	CAAGCAGAAGACGGCATACGAGATggattggGTGACTGGAGTTCAGACGTGT
Sol_iPCR-MPI-110 CCGGCGT	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	CAAGCAGAAGACGGCATACGAGATgcggcaaGTGACTGGAGTTCAGACGTGT
P5 indexing primers (second index)	
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	AATGATACGGCGACCACCGAGATCTACACatcttgcACACTCTTTCCCTACACGACGCTCTT
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	AATGATACGGCGACCACCGAGATCTACACgtaccggACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-11 AACTCCG	AATGATACGGCGACCACCGAGATCTACACcggagttACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-12 TTGAAGT	AATGATACGGCGACCACCGAGATCTACACacttcaaACACTCTTTCCCTACACGACGCTCTT
P5_iPCR-LP-13 ACTATCA	AATGATACGGCGACCACCGAGATCTACACtgatagtACACTCTTTCCCTACACGACGCTCTT
IS4 AGATCTC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
amplification primers for indexed libraries	
IS5	AATGATACGGCGACCACCGA
IS6	CAAGCAGAAGACGGCATACGA
sequencing primer	
P5 index sequencing	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Supplementary Table 2. Index combination used for each library

Experiment MP-CAP

Sample	Index 1	Index 2
L1	AATCTTC (97)	TCGCAGG (1)
L2	ACCAACG (98)	CTCTGCA (2)
L3	AGATGGC (99)	CCTAGGT (3)
L4	CCAGGTT (100)	GGATCAA (4)
L5	CCGTTAG (101)	GCAAGAT (5)
L6	CGCCTCT (102)	ATGGAGA (6)
L7	CTTGCGG (103)	CTCGATG (7)
L8	GGCGGAG (104)	GCTCGAA (8)
L9	TGGACGT (105)	ACCAACT (9)
L10	AACCATG (106)	CCGGTAC (10)
L16	CAGGAAG (107)	AACTCCG (11)
φX	TTGCCGC (control)	AGATCTC (IS4)

Experiments no-CAP and SP-CAP

Sample	Index 1	Index 2	Sample	Index 1	Index 2
L1	AATCTTC (97)	TCGCAGG (1)	L10	AACCATG (106)	CCGGTAC (10)
L2	ACCAACG (98)	CTCTGCA (2)	L11	CCAATCC (109)	ACTATCA (13)
L3	AGATGGC (99)	CCTAGGT (3)	L12	CCGGCGT (110)	TCGCAGG (1)
L4	CCAGGTT (100)	GGATCAA (4)	L13	CGCATAG (111)	CTCTGCA (2)
L5	CCGTTAG (101)	GCAAGAT (5)	L14	CGTAATC (112)	CCTAGGT (3)
L6	CGCCTCT (102)	ATGGAGA (6)	L15	CGTTGGT (113)	GGATCAA (4)
L7	CTTGCGG (103)	CTCGATG (7)	L16	CAGGAAG (107)	AACTCCG (11)
L8	GGCGGAG (104)	GCTCGAA (8)	L17	CTATACG (114)	GCAAGAT (5)
L9	TGGACGT (105)	ACCAACT (9)	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 φX174 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 φX index	Algn Sample index	FSA rate	Algn φX index (QF)	Algn Sample index (QF)	FSA rate (QF)	Kept φX 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%

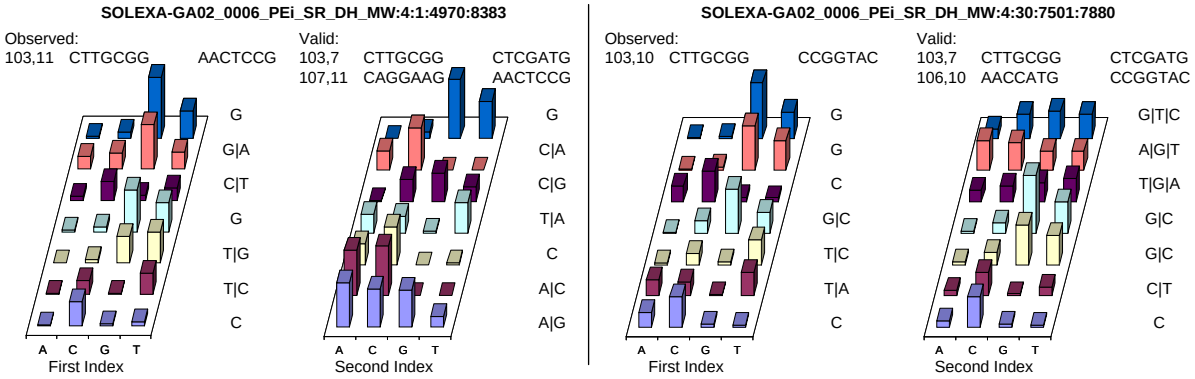
Supplementary Table 5: Quantifying false sample assignments in human and mouse transcriptome sequencing data from single-indexed libraries, which were prepared using Illumina's TruSeq RNA Sample Prep Kit

All single reads were aligned to the φX174 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.

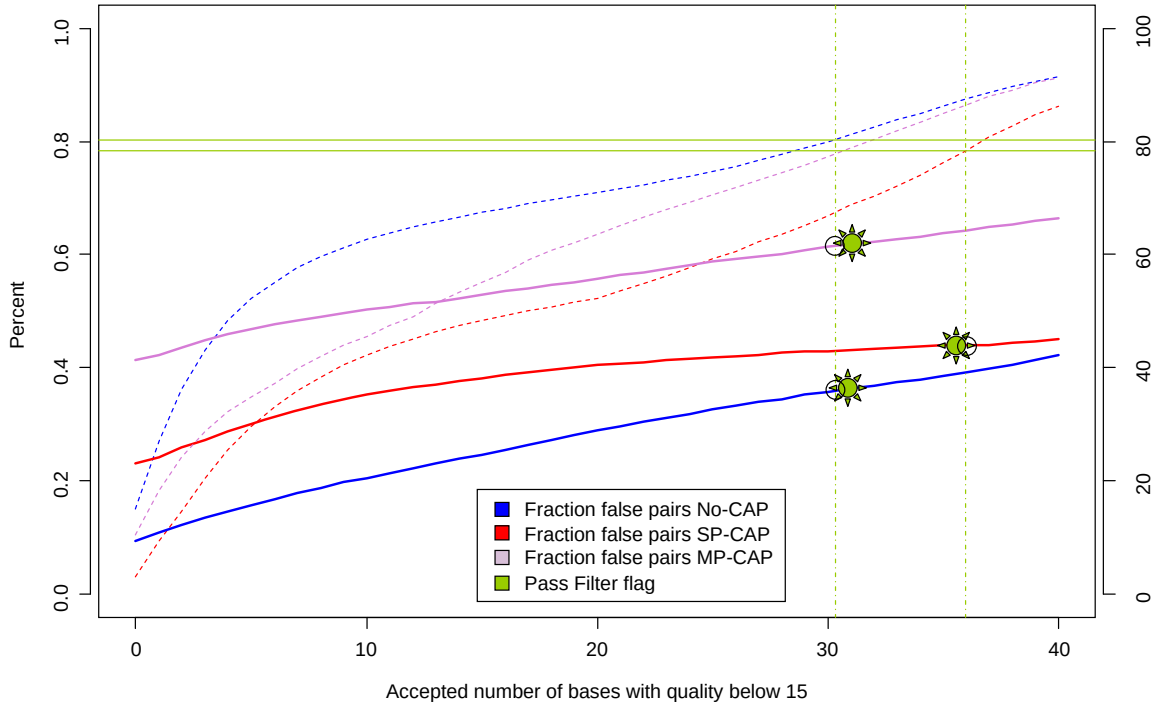
Lane	Raw reads	Algn PhiX index	Ave. algn sample index	False sample assignments	Algn PhiX index QF	Ave. algn sample index QF	False sample assignments QF
1	45990409	563645	906	0.160%	551229	304	0.055%
2	42400537	611873	872	0.142%	600333	307	0.051%
3	44098398	482999	736	0.152%	474908	270	0.057%
4	43447369	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.



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, -]
```

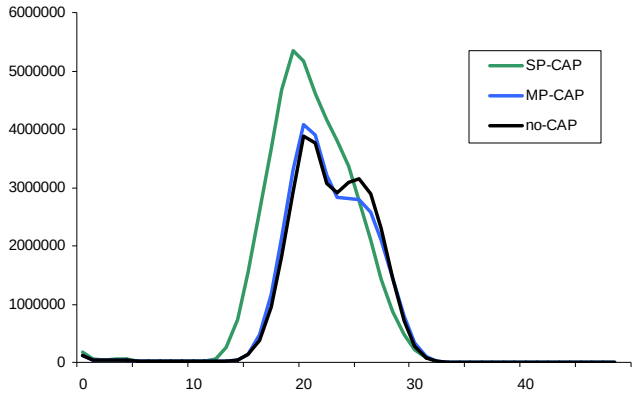
and the 13 indexes of the second index reads:

```
[-, 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, 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 revere 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	GGCGGAG,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					103,9	CTTGCGG,ACCAACT			
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					103,11	CTTGCGG,AACTCCG			
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					107,9	CAGGAAG,ACCAACT			
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					107,9	CAGGAAG,ACCAACT			
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					105,13	TGGACGT,ACTATCA			
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					104,13	GGCGGAG,ACTATCA			
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					106,9	AACCATG,ACCAACT			
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 (11.499E-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 (6.855E-5) the indexes read out from the second index did not agree. We obtained similar rates for SP-CAP (Index1: 209/1,699,585 = 12.297E-5, Index2: 138/1,699,585 = 8.120E-5) and MP-CAP (Index1: 305/3,451,555 = 8.837E-5, Index2: 275/3,451,555 = 7.967E-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

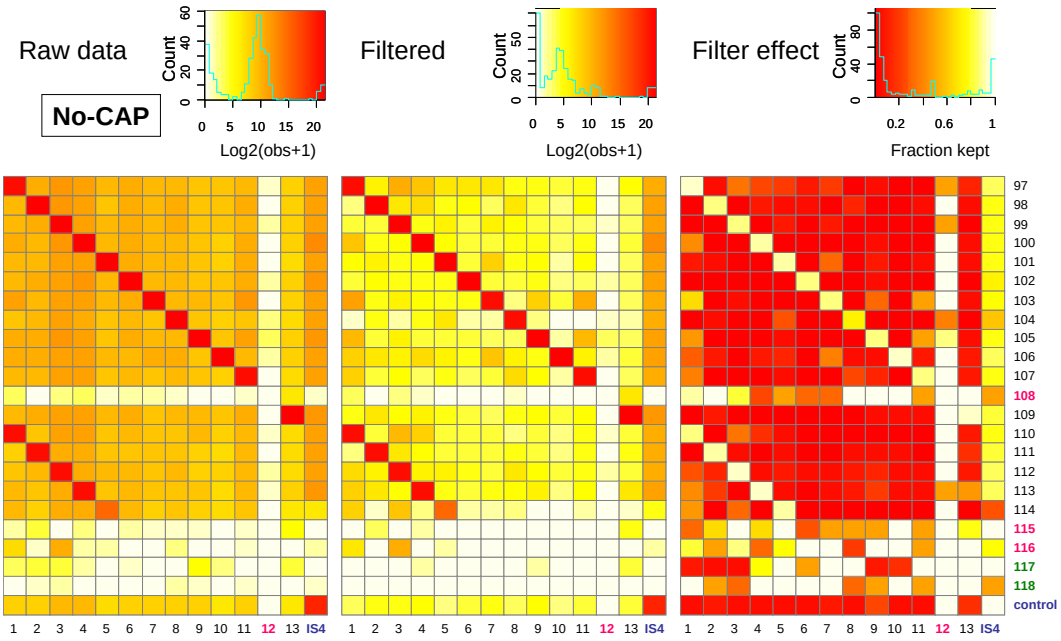
We checked the frequency of all putative index pairs, also including (i) index sequences that have not participated in the experiment at all, but were synthesized in the same batch of indexed oligonucleotides, and (ii) indexes that participated in some step of the experiment but did not participate in 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.

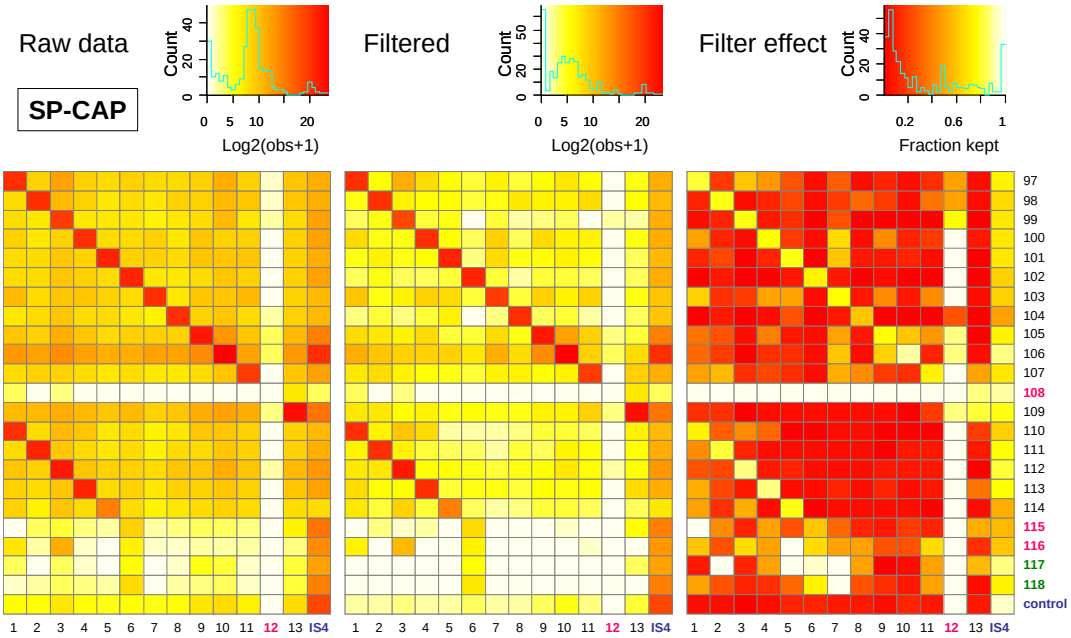
11/103 and 11/105 in no-CAP, 97/3 and 10/105 in SP-CAP, and 97/3 in MP-CAP that are overrepresented compared to background.

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).

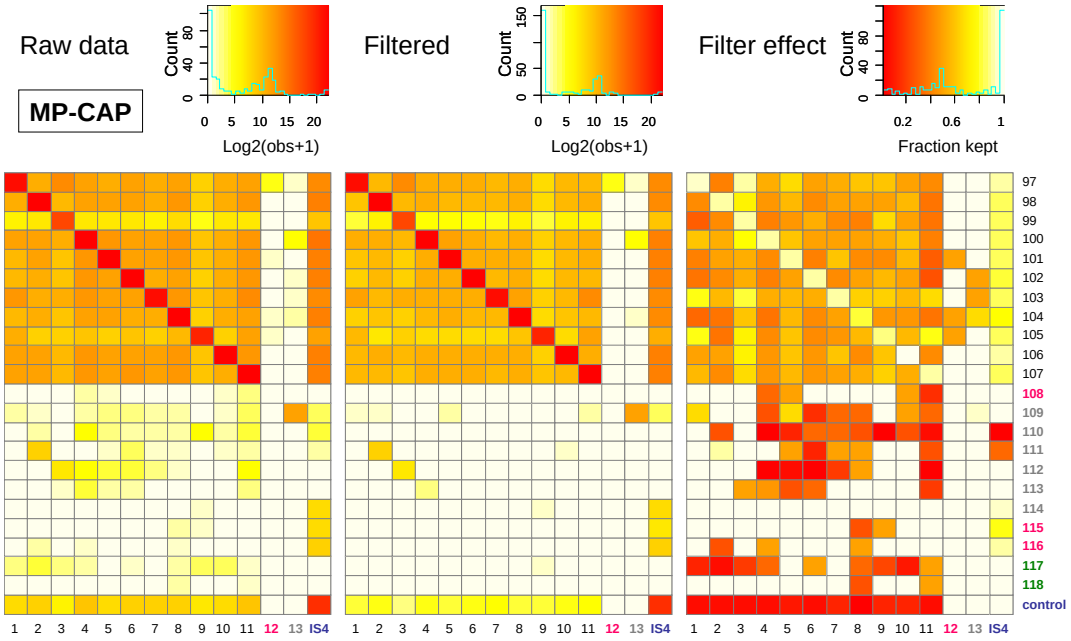
Supplementary Figure 4: Counts of all index pairs before and after applying a minimum base quality score cutoff of 15 to the index reads. Shown are the result for experiment no-CAP, where libraries were amplified independently and pooled just prior to sequencing. The first index is plotted on the vertical 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 sequencing.



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	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:

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

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 φX174 reference genome using BWA(4) and identified false index assignments as reads with a φX 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% φX174 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.