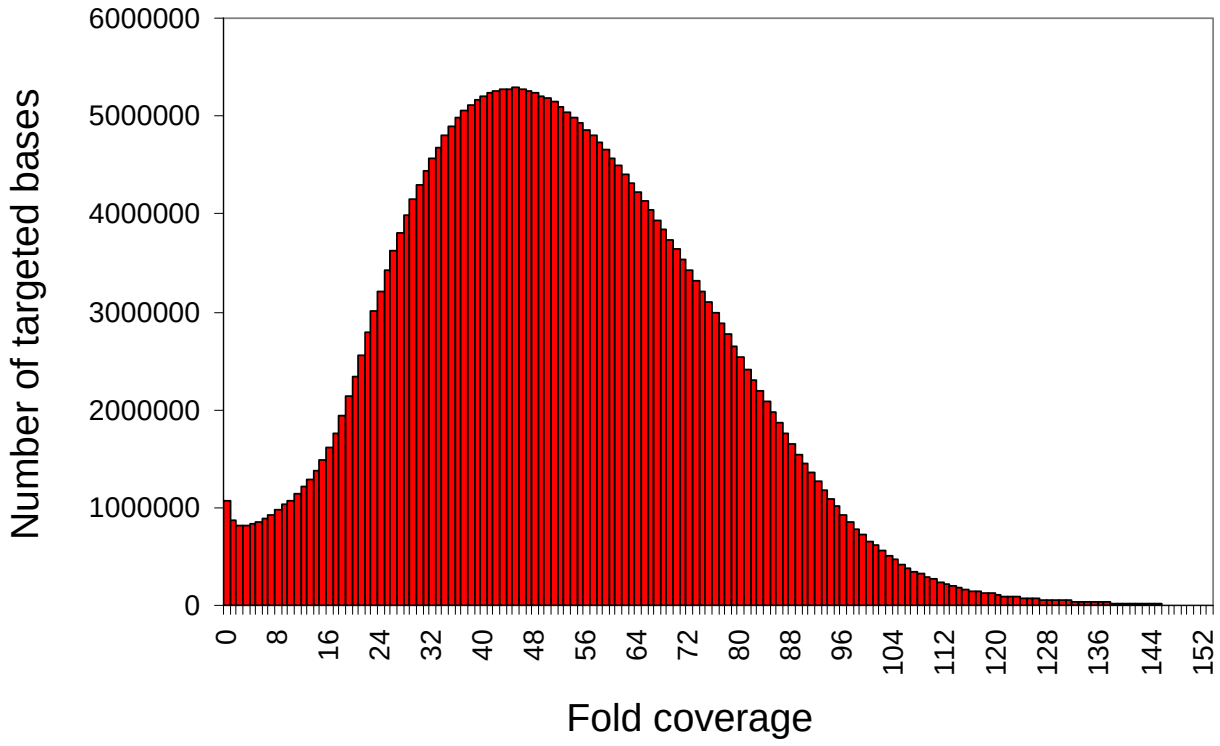
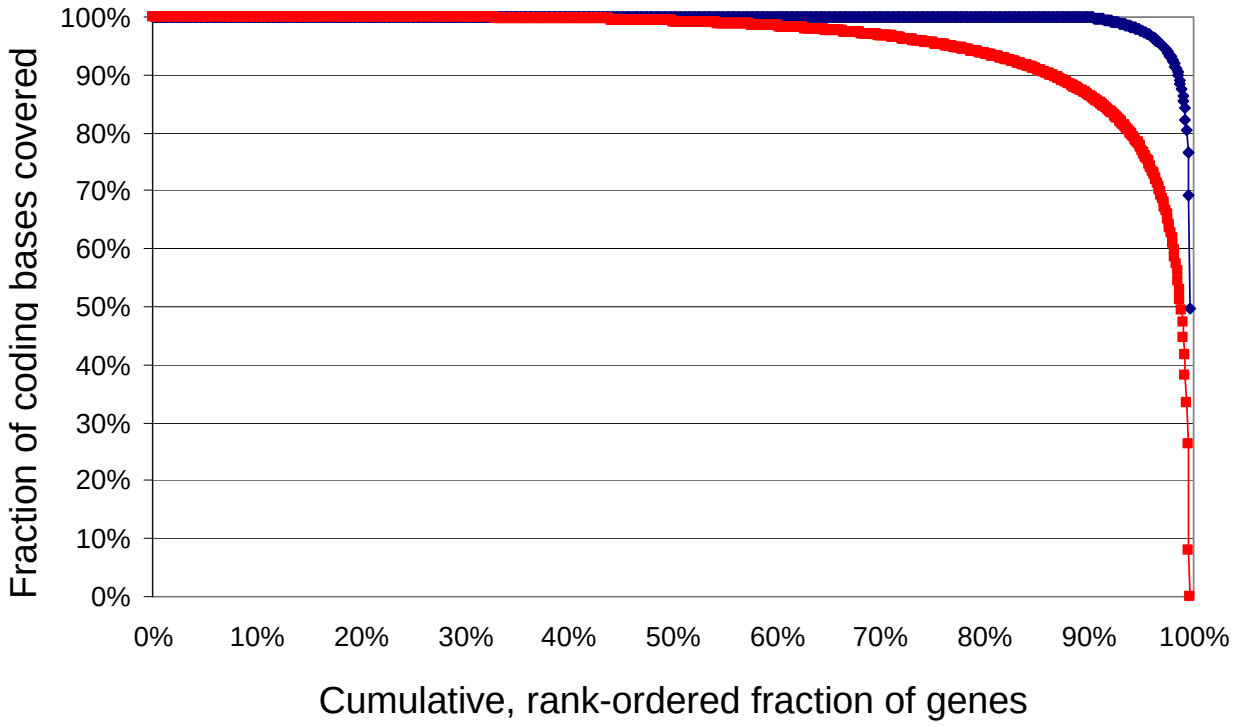
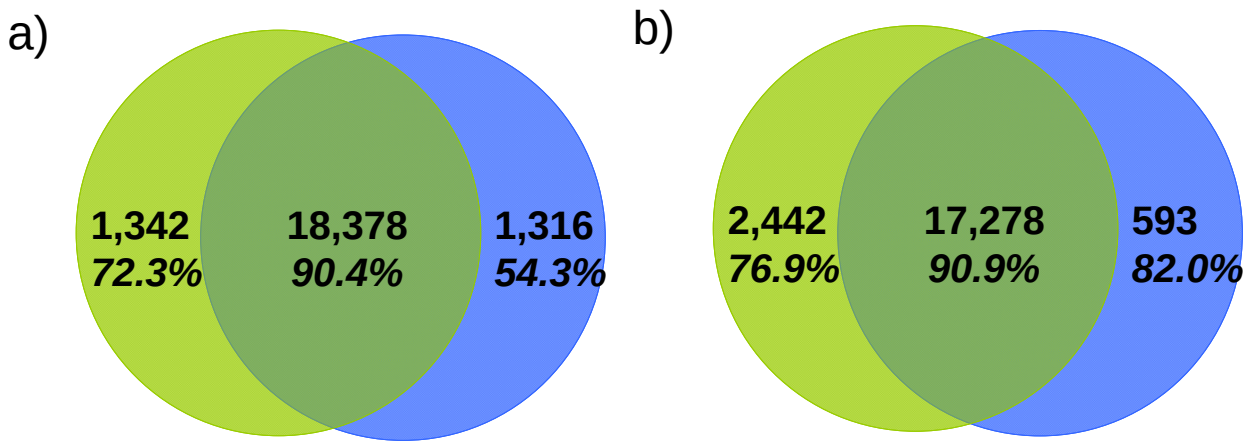


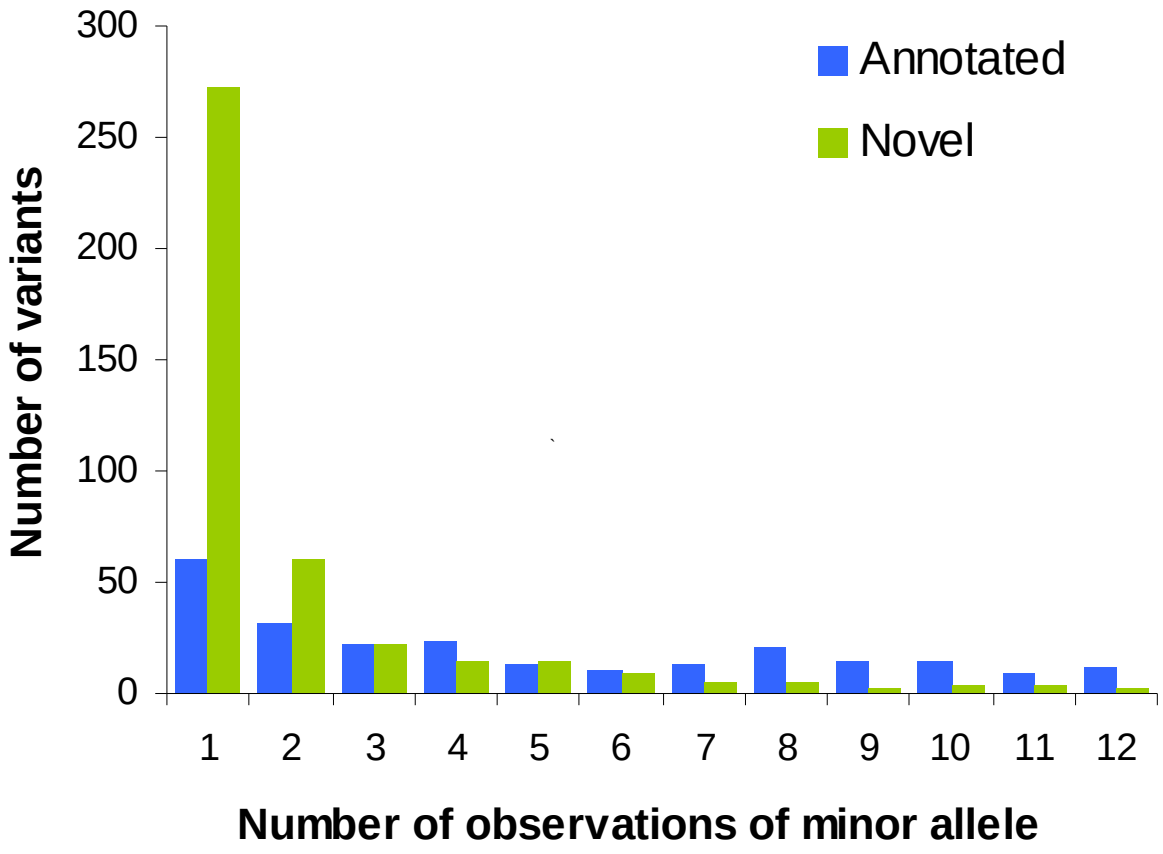
Supplementary Figure 1. Non-cumulative histogram of fold-coverage across 12 exomes. The distribution of fold-coverage of mappable, targeted bases (26.6 Mb), summed across the twelve exome datasets (318 Mb aggregate target), is shown. As potential PCR duplicates (reads with the same start-point and orientation within the same genomic library) have been filtered out, the maximum possible coverage of any given position is 152x (i.e. reads from only 76 potential start-points x 2 orientations).



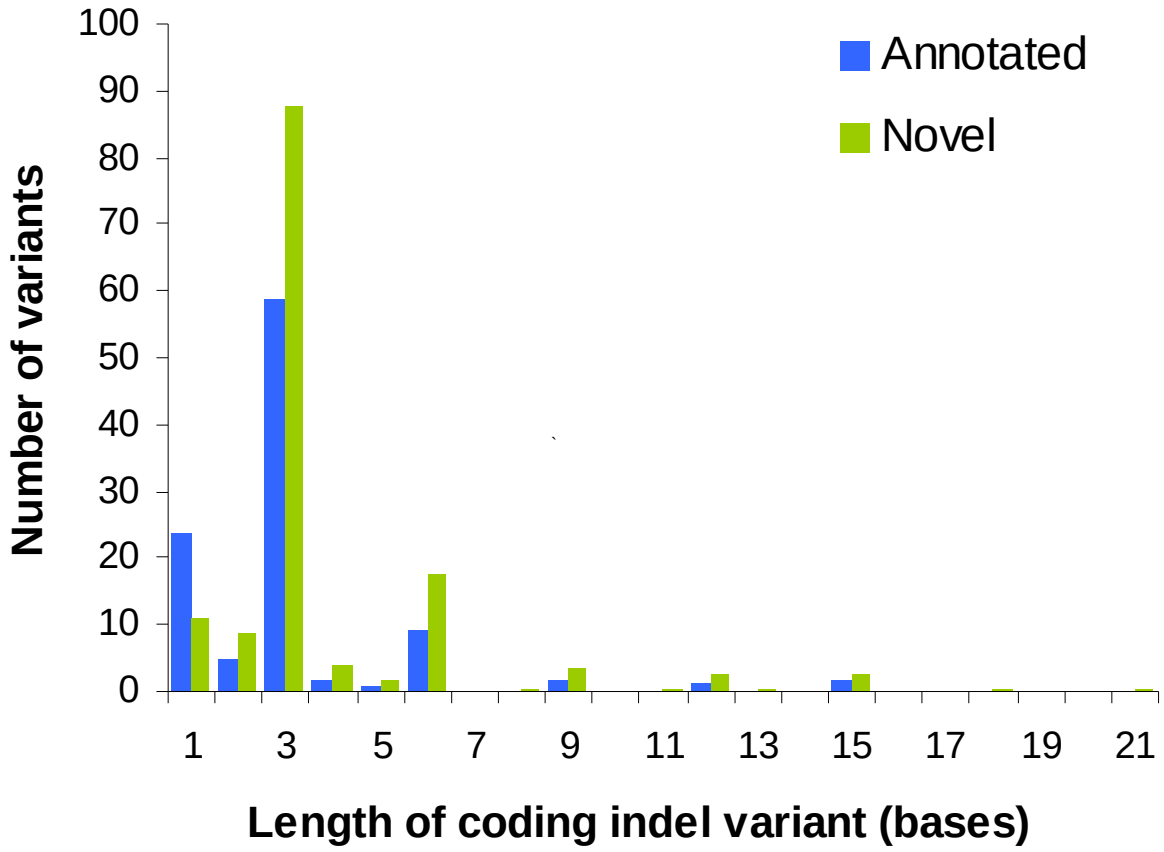
Supplementary Figure 2. Distribution of completeness on a gene-by-gene basis. We calculated the fraction of coding bases covered at least 1x (blue) or with sufficient coverage to variant call (red) on a gene-by-gene basis, for 197,952 genes (16,496 genes x 12 individuals). The cumulative, rank-ordered fraction of genes that meet minimum criteria are plotted above. For example, 98% of genes >=1x coverage for at least 95% of their coding bases, while 78% of genes had sufficient coverage for variant calling for at least 95% of their coding bases.



Supplementary Figure 3. Comparison of cSNPs from exome sequencing and whole genome sequencing of NA18507. We identified 19,720 cSNPs by exome sequencing of this individual (green circles). In the above Venn diagrams, these are intersected with cSNPs from whole genome sequencing of this individual that overlapped with our exome target (Bentley et al. (2008)) (blue circles). The percentage of each subset confirmed by dbSNP is given below the count. In (a), we used all cSNPs from Bentley et al. (2008) that were called by Maq (n = 19,694), whereas in (b), we used only the high-confidence cSNPs from Bentley et al. (2008), i.e. called by both Maq and Eland (n = 17,871). In (a), we observe similar numbers of variants that are only called in one data-set, suggesting both sets contain false negatives. The dbSNP-based confirmation rate is modestly higher for the 1,342 variants that are only called in our data-set as compared to the 1,316 variants only called from Bentley et al. (2008) (72.3% versus 54.3%). In (b), we call a much larger number of cSNPs than the high-confidence set from whole genome sequencing (19,720 vs. 17,871). The dbSNP-based confirmation rates for cSNPs called in only one dataset are similar (76.9% versus 82.0%). 78% of the 593 cSNPs called by Bentley et al. (2008) (Maq-Eland intersection) that were not identified here were at coordinates that had insufficient coverage to call in our data.



Supplementary Figure 4. Minor allele frequencies for novel versus previously annotated coding indels. A histogram of the number of observations of the minor allele for all observed coding indels is shown. Previously annotated (blue) and novel (green) coding indels are plotted separately.



Supplementary Figure 5. Distribution of lengths of novel versus previously annotated coding indels. The average number of coding indels called in each of the 12 exomes with various lengths is shown, plotted separately by annotation status.

# genes in which each affected has at least one…	One Affected	Two Affecteds	Three Affecteds	All Four Affected
nonsynonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510 – 4,617	3,284 – 3,373	2,765 – 2,808	2,479
NS/SS/I not in dbSNP	513 – 603	115 – 131	67 – 71	53
NS/SS/I not in 8 HapMap exomes	799 – 912	150 – 191	49 – 54	21
NS/SS/I neither in dbSNP nor 8 HapMap exomes	360 – 435	29 – 38	4 – 8	1 (MYH3)
… AND predicted to be damaging	160 - 210	5 – 14	1 – 2	1 (MYH3)

Supplementary Figure 6. Direct identification of the causal gene for a monogenic disorder by exome sequencing. Each box lists the number of genes with 1+ nonsynonymous cSNP, splice-site variant, or coding indel ("NS/SS/I" variants) meeting specific criteria. In columns, we show the effect of requiring that 1+ NS/SS/I variants be observed in a given gene in all of 1, 2, 3 or 4 FSS-affected individuals. This figure is identical in format to **Figure 2,** except that we here provide ranges of observations that occur when all possible permutations of 1, 2, or 3 FSS-affected individuals are used.

Individual	Mapped Bases	On Target Bases	Near Target Bases	Capture Specificity
NA18507 (YRI)	6,607,484,688	2,888,661,709	661,786,591	54%
NA18517 (YRI)	6,494,342,272	2,545,603,815	595,560,089	48%
NA19129 (YRI)	6,297,755,808	2,643,637,650	580,152,618	51%
NA19240 (YRI)	5,986,557,544	2,680,726,944	570,797,168	54%
NA18555 (CHB)	6,006,434,128	2,367,312,059	533,865,581	48%
NA18956 (JPT)	6,696,487,148	2,872,329,417	681,969,727	53%
NA12156 (CEU)	5,807,479,732	1,776,298,679	389,939,201	37%
NA12878 (CEU)	7,412,509,748	3,006,930,065	691,105,599	50%
FSS10066 (Eur)	6,213,695,628	2,724,817,939	522,381,165	52%
FSS10208 (Eur)	6,828,499,072	2,779,965,715	545,984,133	49%
FSS22194 (Eur)	6,710,279,780	2,139,816,034	523,364,262	40%
FSS24895 (Eur)	5,806,226,492	2,472,076,217	528,472,867	52%
Average	**6,405,646,003**	**2,574,848,020**	**568,781,583**	**49%**

Supplementary Table 1. Sequencing of twelve exome-enriched shotgun libraries. Summary statistics on massively parallel sequencing are shown. All data was collected as unpaired 76 bp reads on the Illumina Genome Analyzer II platform (~10 lanes per individual). For each individual, we show the total number of mapped bases (Maq mapping score > 0), the number of these that align within ("On Target Bases") or near ("Near Target Bases") the 164,007 targeted intervals. Near target bases do not fall within a target, but are from reads that directly overlap a target. Capture specificity is calculated as the fraction of reads overlapping a target.

	Albert *et al.*	Hodges *et al.*	This study
% of reads mapping to exon target	36%-76%	36%-55%	37%-54%
% of target bases covered >=1x	91.3-98.2%	25%	99.5-99.8%
Estimated sensitivity for variant calling	62.9-87.8%	60%*	95.5-97.1%

Supplementary Table 2. Comparison to past reports on exonic or exomic array-based capture. For Albert *et al*., metrics for capture of 6,726 exonic regions are taken from Supplementary Tables 1 & 2, and we assume that at least 2 non-reference observations would be required to call a variant. For Hodges *et al*., metrics for all-exome capture are taken from the text or Table 1. *reported for one of seven arrays only (EC5)

Individual	# of coding indels	% in dbSNP	% heterozygous	% 3n in length	% insertions
NA18507 (YRI)	189	60%	61%	68%	40%
NA18517 (YRI)	204	58%	68%	70%	41%
NA19129 (YRI)	196	56%	65%	64%	42%
NA19240 (YRI)	183	57%	67%	70%	38%
NA18555 (CHB)	145	76%	53%	71%	47%
NA18956 (JPT)	139	71%	56%	71%	47%
NA12156 (CEU)	163	66%	61%	75%	48%
NA12878 (CEU)	146	64%	58%	66%	47%
FSS10066 (Eur)	170	59%	61%	70%	49%
FSS10208 (Eur)	165	59%	64%	70%	44%
FSS22194 (Eur)	152	66%	66%	68%	45%
FSS24895 (Eur)	142	65%	51%	65%	44%
Average	**166**	**63%**	**61%**	**69%**	**44%**

Supplementary Table 3. Coding indels across 12 human exomes. The number of called indels in each individual is listed, along with the proportion of these that are annotated in dbSNP (v129), the proportion that are heterozygous, the proportion that have a length that is a multiple of 3, and the proportion that are insertions as opposed to deletions, relative to the reference genome (hg18). YRI = Yoruba HapMap; CHB = Chinese HapMap; JPT = Japanese HapMap; CEU = CEPH HapMap; Eur = European-American ancestry (non-HapMap).

Individual	Synonymous cSNPs	Missense cSNPs	Nonsense cSNPs	Splice-site SNPs	In-frame Indel	Frameshift Indel
NA18507 (YRI)	10817 (987,1620)	8862 (1143,1629)	41 (13,17)	18 (5,8)	128 (48,23)	61 (28,19)
NA18517 (YRI)	10845 (1111,1693)	8838 (1282,1644)	54 (18,18)	19 (6,9)	142 (60,24)	62 (26,16)
NA19129 (YRI)	10950 (1190,1699)	8761 (1258,1621)	50 (15,18)	18 (7,5)	125 (54,26)	71 (32,17)
NA19240 (YRI)	10749 (1109,1637)	8719 (1225,1582)	49 (15,16)	18 (8,7)	128 (56,22)	55 (23,15)
NA18555 (CHB)	8807 (489,653)	7198 (650,790)	42 (14,8)	21 (5,11)	103 (21,7)	42 (14,7)
NA18956 (JPT)	8706 (507,674)	7257 (643,785)	48 (13,11)	18 (1,3)	98 (30,13)	41 (11,8)
NA12156 (CEU)	8750 (355,540)	7322 (503,712)	47 (11,5)	15 (3,2)	123 (46,17)	40 (10,7)
NA12878 (CEU)	8706 (380,516)	7225 (529,632)	39 (10,10)	18 (8,10)	97 (33,13)	49 (20,5)
FSS10066 (Eur)	8822 (441,564)	7362 (629,703)	45 (15,12)	14 (2,2)	119 (49,18)	51 (21,12)
FSS10208 (Eur)	8709 (425,520)	7322 (617,668)	42 (13,8)	16 (2,6)	116 (48,18)	49 (20,12)
FSS22194 (Eur)	8745 (374,493)	7298 (580,652)	51 (12,13)	14 (2,3)	104 (34,10)	48 (18,6)
FSS24895 (Eur)	8783 (393,518)	7157 (553,579)	46 (13,8)	12 (6,5)	92 (31,7)	50 (19,9)
Average (YRI)	10840 (1099,1662)	8795 (1227,1619)	49 (15,17)	18 (7,7)	131 (55,24)	62 (27,17)
Average (Non-YRI)	8754 (421,560)	7268 (588,690)	45 (13,9)	16 (4,5)	107 (37,13)	46 (17,8)
Average	9449 (647,927)	7777 (801,1000)	46 (14,12)	17 (5,6)	115 (43,17)	52 (20,11)

Supplementary Table 4. Numbers of variants observed in each individual. A list of the number of observations of synonymous cSNPs, missense cSNPs, nonsense cSNPs, SNPs that disrupt canonical splice-site bases, in-frame coding indels, and frameshift indels. The first and second numbers in parentheses refer to novel variants (i.e. not in dbSNP) and singleton observations, respectively. Averages (Yoruba, non-African, and overall) are also shown.

Oligonucleotide	Sequence	Function
SLXA_1_HI	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT	Library Adaptor
SLXA_1_LO	/5Phos/GAT CGG AAG AGC GTC GTG TAG GGA AAG AGT GT	Library Adaptor
SLXA_2_HI	CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T	Library Adaptor
SLXA_2_LO	/5Phos/GAT CGG AAG AGC TCG TAT GCC GTC TTC TGC TTG	Library Adaptor
SLXA_FOR_AMP	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T	Library Amplification, Hybridization Blocker
SLXA_REV_AMP	CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T	Library Amplification, Hybridization Blocker
SLXA_REV_AMP_rev	AGA TCG GAA GAG CTC GTA TGC CGT CTT CTG CTT G	Hybridization Blocker
SLXA_FOR_AMP_rev	AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AGA TCT CGG TGG TCG CCG TAT CAT T	Hybridization Blocker

Supplementary Table 5. Sequences of oligonucleotides used for library construction or blocking. Oligonucleotide sequences