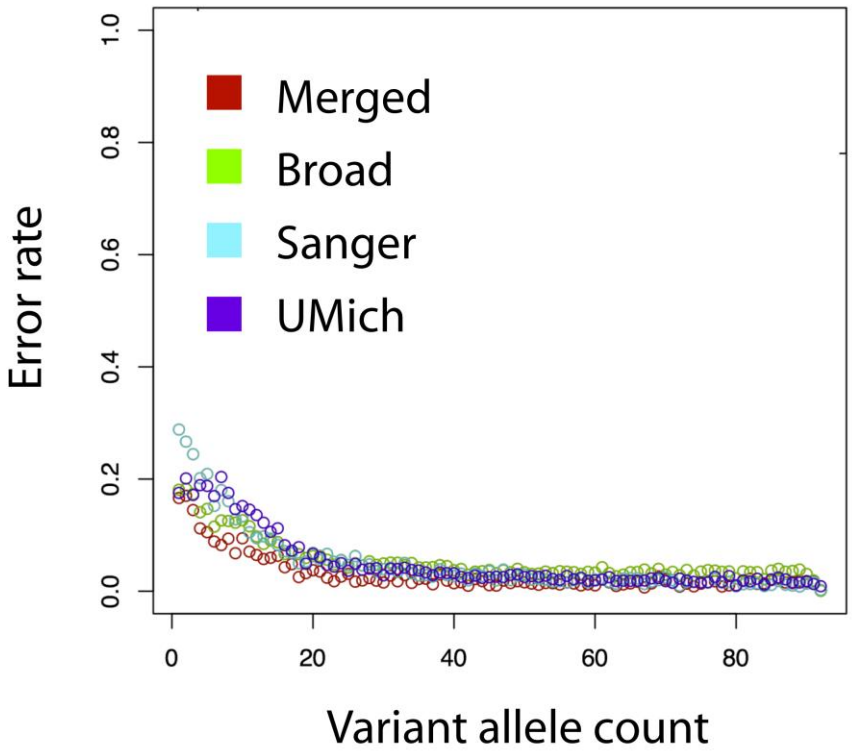
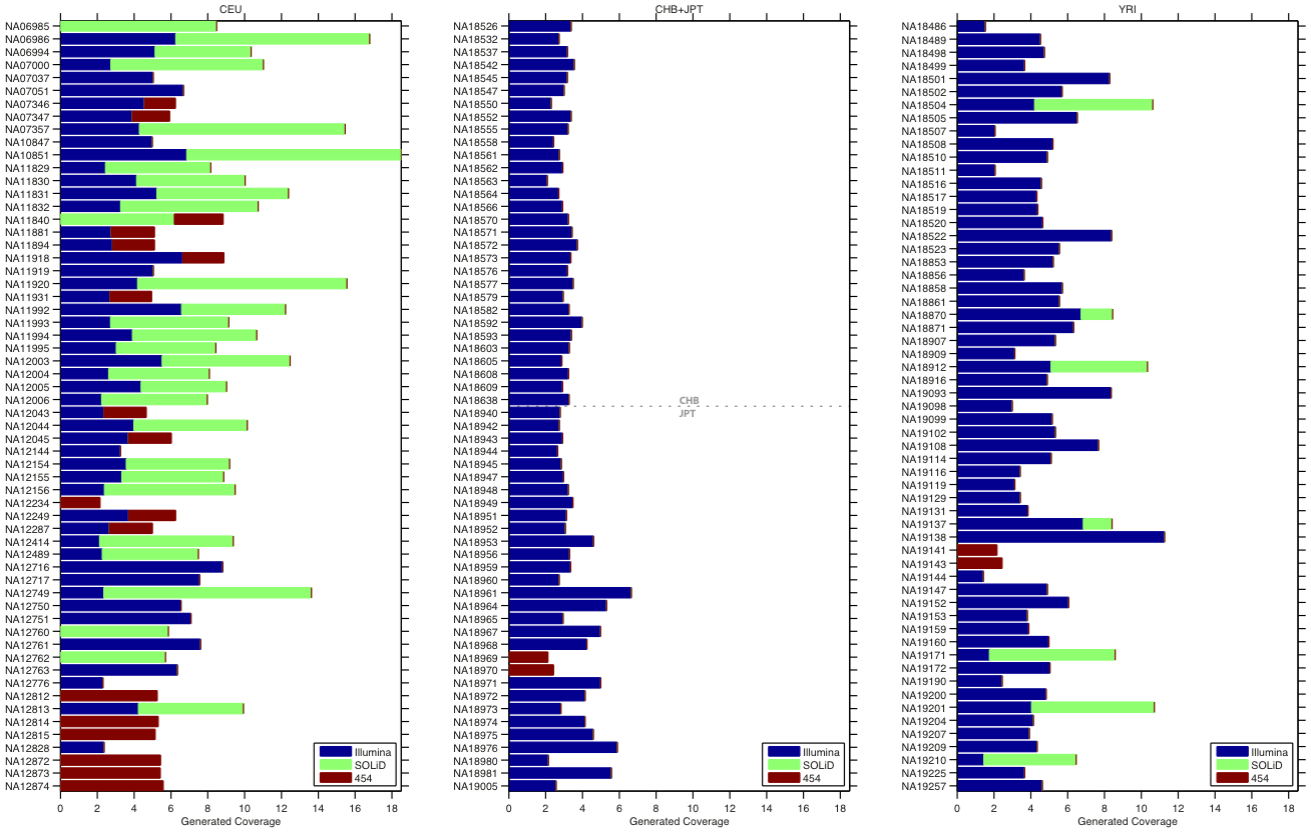
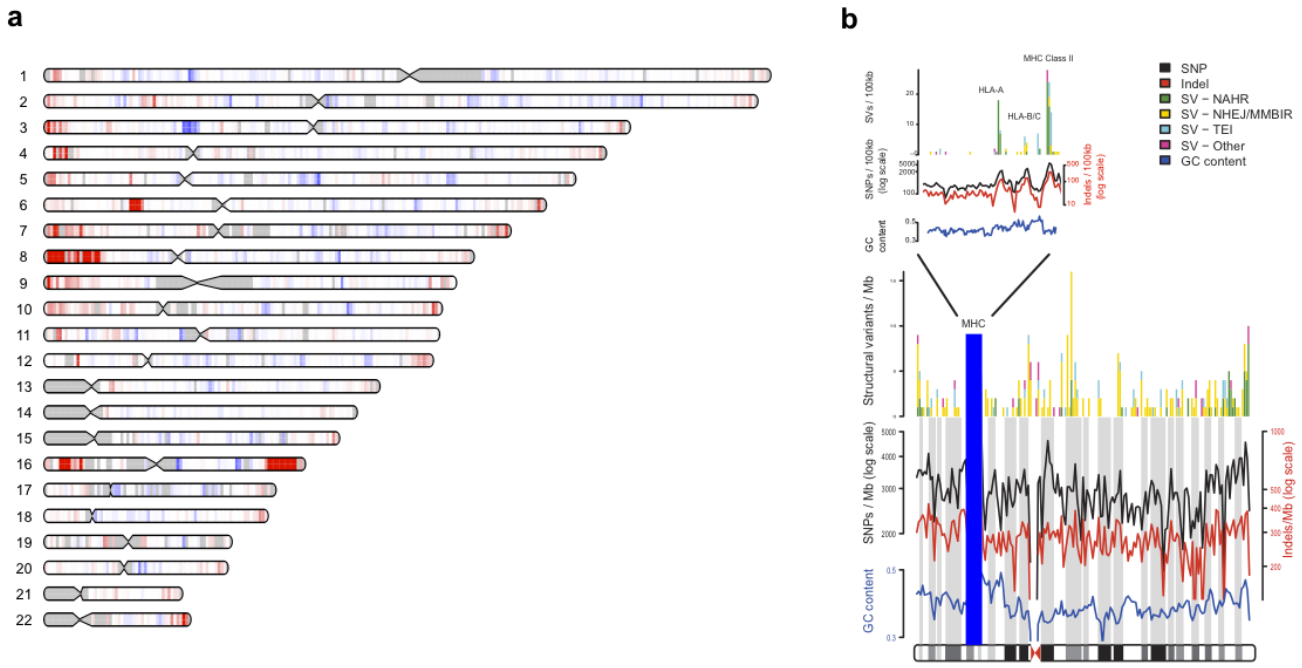
Supplementary Figures



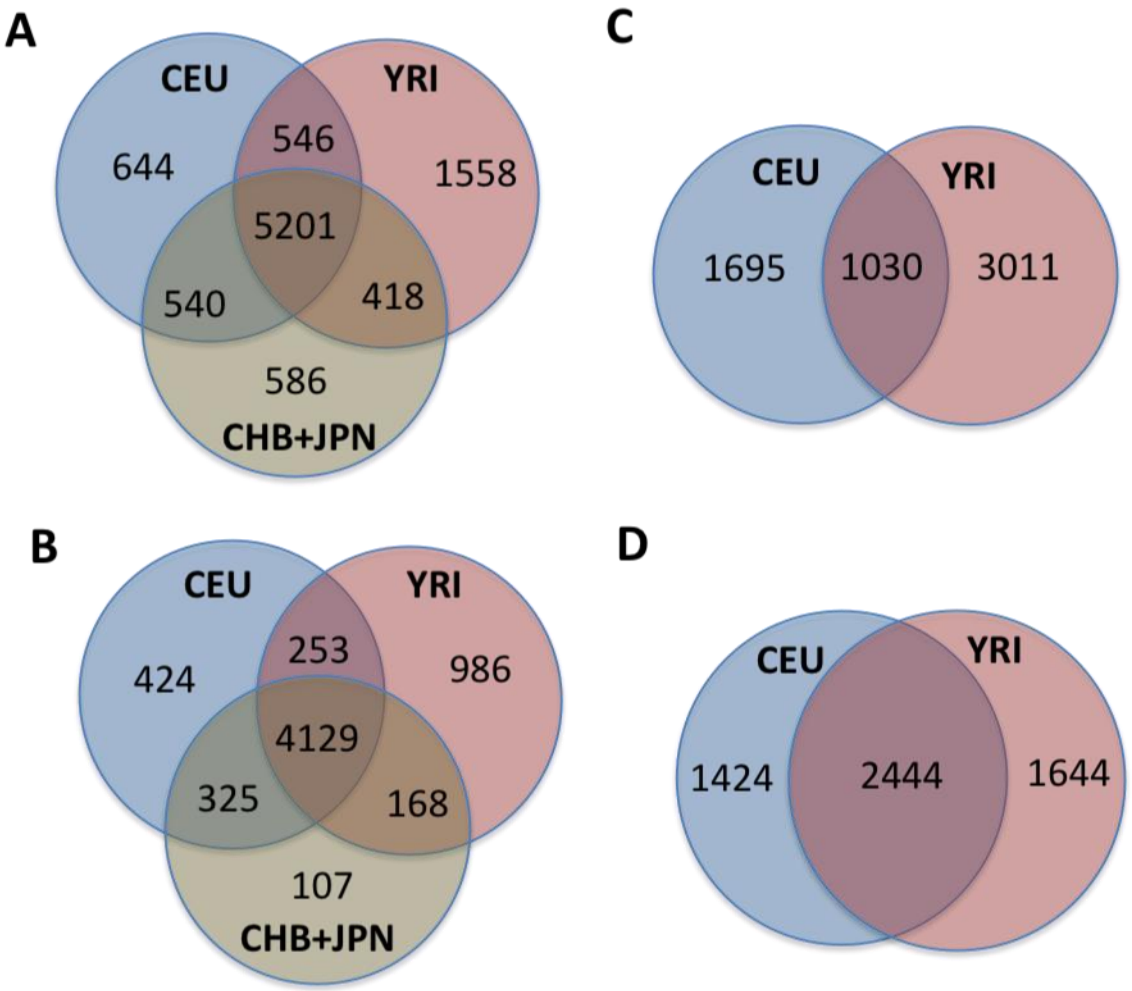
Supplementary Figure 1 – Genotype accuracy of SNP call sets, compared to 1000G genotyping chip. The SNP call sets from each centre (Broad – green, Sanger – cyan, UMich – purple) have comparable performance. However, the merged call set (red), which is based on consensus genotypes, has consistently higher accuracy.



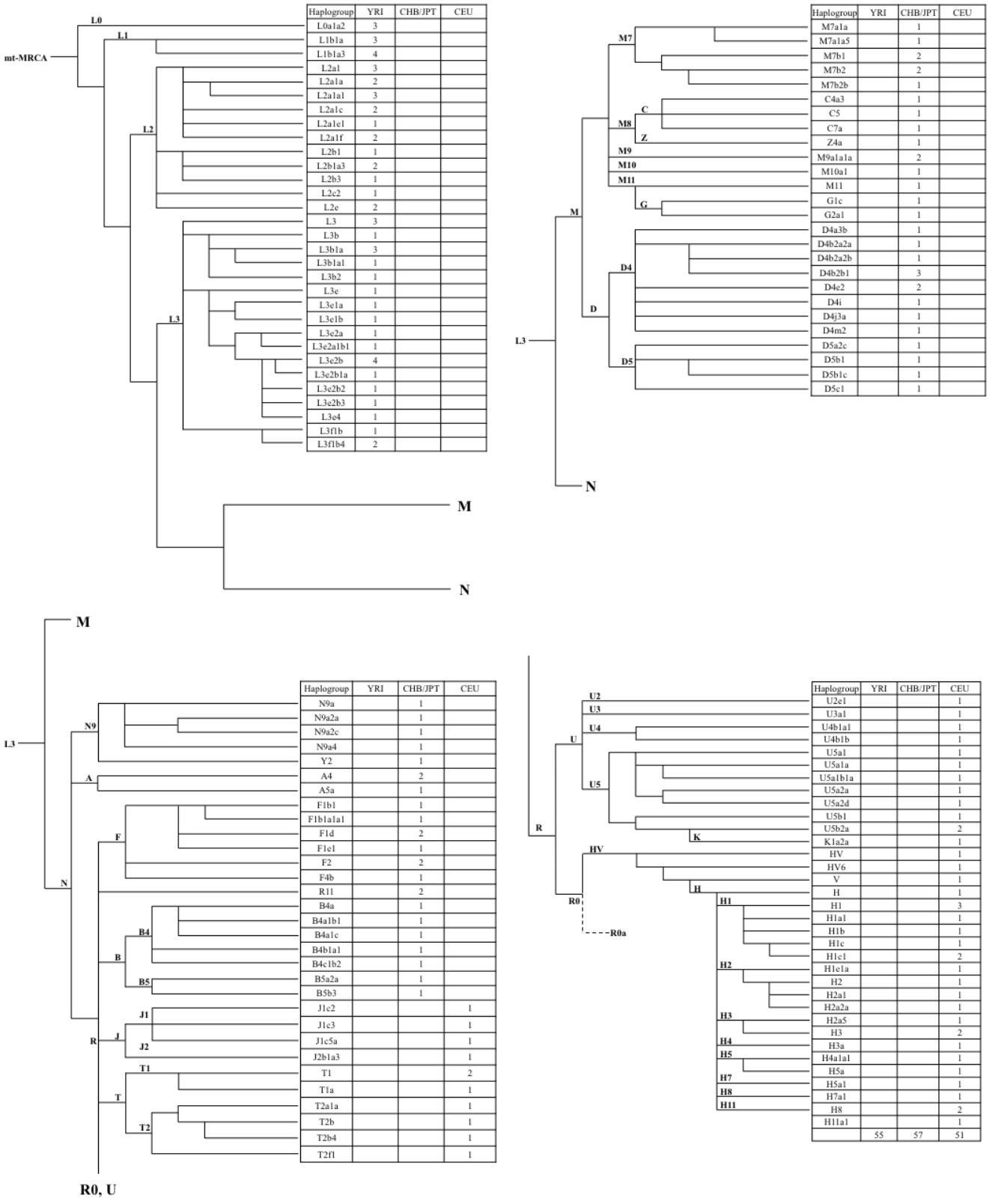
Supplementary Figure 2 - Generated sequence by individual. This figure shows the amount of generated sequence for each individual in the Low Coverage Pilot, broken down by sequencing technology; Dark Blue = Illumina, Green = SOLiD, Dark Red = 454.



Supplementary Figure 3 - Variation across the genome. (a) Genomic distribution of SNP density on the autosomes. The colours show the SNP density in 1 Mb bins, with red indicating higher densities and blue indicating lower densities(Kin and Ono 2007). SNP densities were calculated as the number of SNPs divided by the number of callable bases in 1 Mb bins. Bins for which less than 75% of bases were callable are shown in grey. The colours cover the median SNP density +- 2.58 standard deviations. Note high rates of SNP variation at the HLA on 6p and sub-telomeric regions and a 5Mb region of very low diversity on 3p21. (b) Distribution of variants on chromosome 6 in the low coverage pilot. From the bottom upwards are shown GC content (blue), the density of SNPs (black) and small indels (red) in the CEU population, and the 1017 structural variants (SV) on chromosome 6 classified by type (NAHR: Non-Allelic Homologous Recombination, NHEJ/MMBIR: Non Homologous End-Joining/Microhomology Mediated Break Induced Replication, TEI: Transposable Element Insertion, or Other). The MHC region is inset with different axes to reflect the greatly increased diversity there. Bin sizes are 1 Mb in the whole chromosome region, and 100 kb in the MHC.



Supplementary Figure 4 - Population origin of known and existing deletions in the SV discovery set (A) Novel deletions discovered in low-coverage sequencing data. (B) Previously known deletions, discovered in low-coverage sequencing data. (C) Novel deletions discovered in trio sequencing data. (D) Previously known deletions, discovered in trio sequencing data.



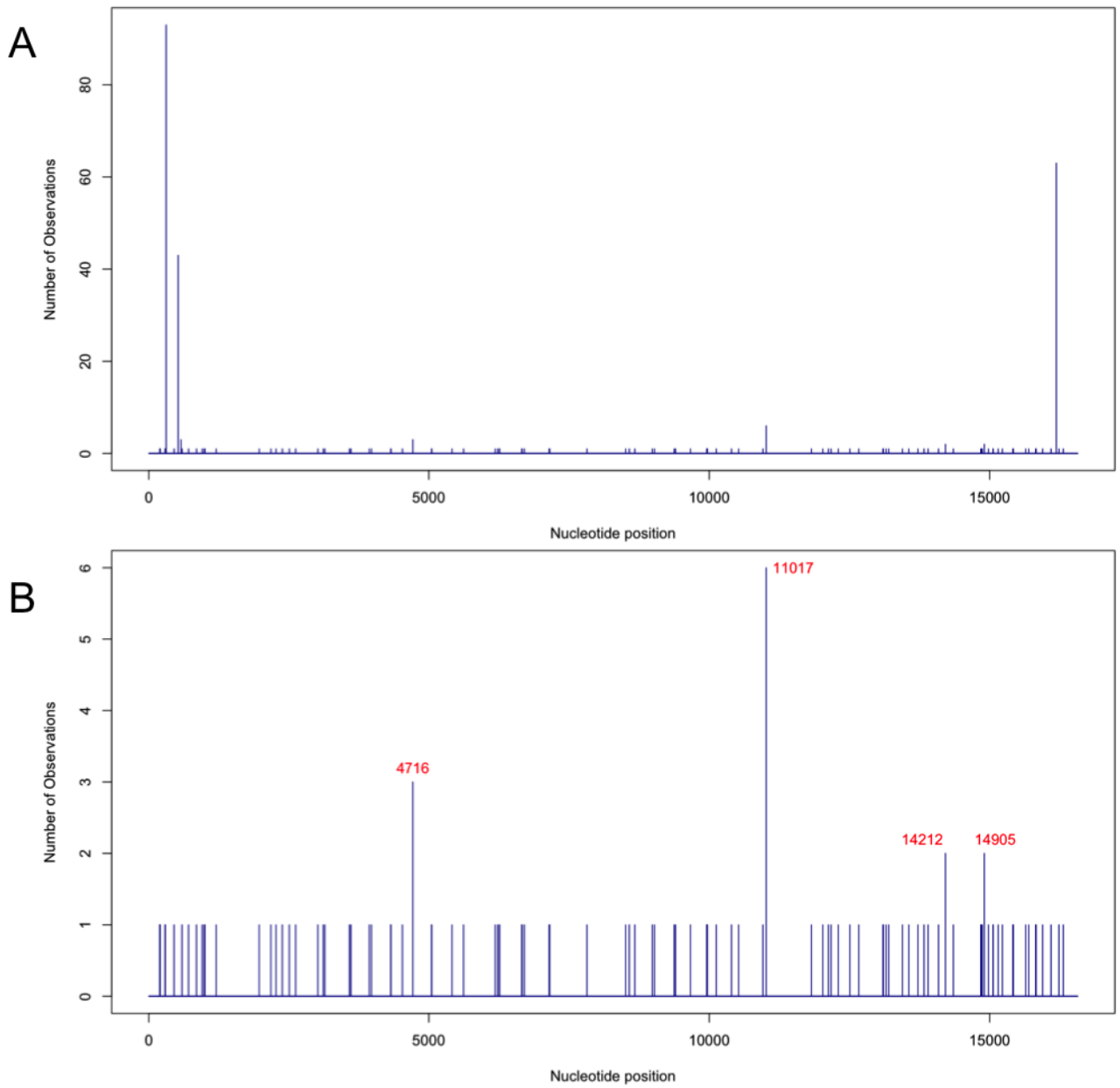
Haplogroup	YRI	CHB/JPT	CEU
L0a1a2	3		
L1b1a	3		
L1b1a3	4		
L2a1	3		
L2a1a	2		
L2a1a1	3		
L2a1c	2		
L2a1e1	1		
L2a1f	2		
L2b1	1		
L2b1a3	2		
L2b3	1		
L2c2	1		
L2e	2		
L3	3		
L3b	1		
L3b1a	3		
L3b1a1	1		
L3b2	1		
L3e	1		
L3e1a	1		
L3e1b	1		
L3e2a	1		
L3e2a1b1	1		
L3e2b	4		
L3e2b1a	1		
L3e2b2	1		
L3e2b3	1		
L3e4	1		
L3f1b	1		
L3f1b4	2		

Haplogroup	YRI	CHB/JPT	CEU
M7a1a		1	
M7a1a5		1	
M7b1		2	
M7b2		2	
M7b2b		1	
C4a3		1	
C5		1	
C7a		1	
Z4a		1	
M9a1a1a		2	
M10a1		1	
M11		1	
G1c		1	
G2a1		1	
D4a3b		1	
D4b2a2a		1	
D4b2a2b		1	
D4b2b1		3	
D4e2		2	
D4i		1	
D4j3a		1	
D4m2		1	
D5a2c		1	
D5b1		1	
D5b1c		1	
D5c1		1	

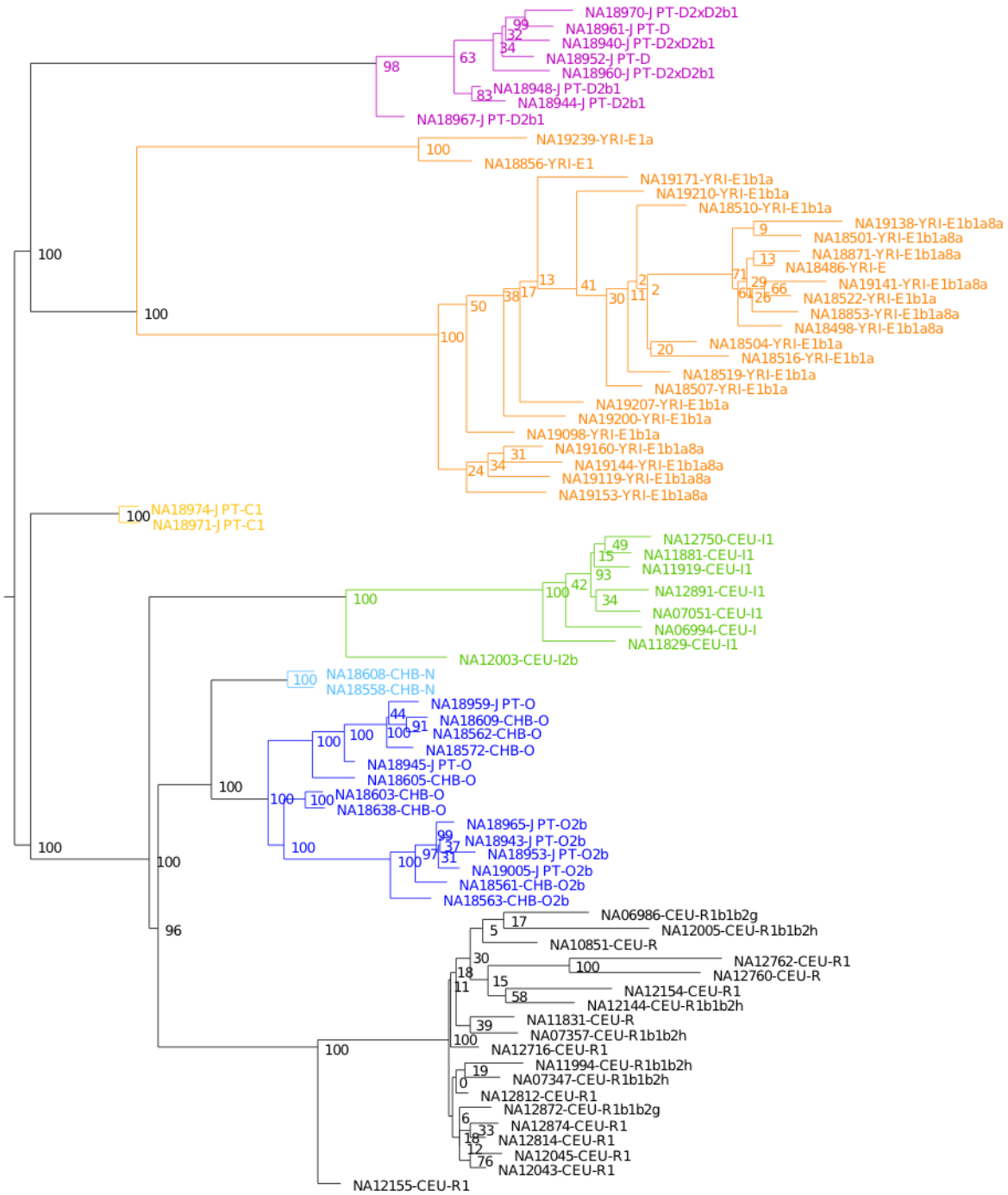
Haplogroup	YRI	CHB/JPT	CEU
N9a		1	
N9a2a		1	
N9a2c		1	
N9a4		1	
Y2		1	
A4		2	
A5a		1	
F1b1		1	
F1b1a1a1		1	
F1d		2	
F1e1		1	
F2		2	
F4b		1	
R11		2	
B4a		1	
B4a1b1		1	
B4a1c		1	
B4b1a1		1	
B4c1b2		1	
B5a2a		1	
B5b3		1	
J1c2			1
J1c3			1
J1c5a			1
J2b1a3			1
T1			2
T1a			1
T2a1a			1
T2b			1
T2b4			1
T2f1			1

Haplogroup	YRI	CHB/JPT	CEU
U2e1			1
U3a1			1
U4b1a1			1
U4b1b			1
U5a1			1
U5a1a			1
U5a1b1a			1
U5a2a			1
U5a2d			1
U5b1			1
U5b2a			2
K1a2a			1
HV			1
HV6			1
V			1
H			1
H1			3
H1a1			1
H1b			1
H1c			1
H1c1			2
H1e1a			1
H2			1
H2a1			1
H2a2a			1
H2a5			1
H3			2
H3a			1
H4a1a1			1
H5a			1
H5a1			1
H7a1			1
H8			2
H11a1			1
	55	57	51

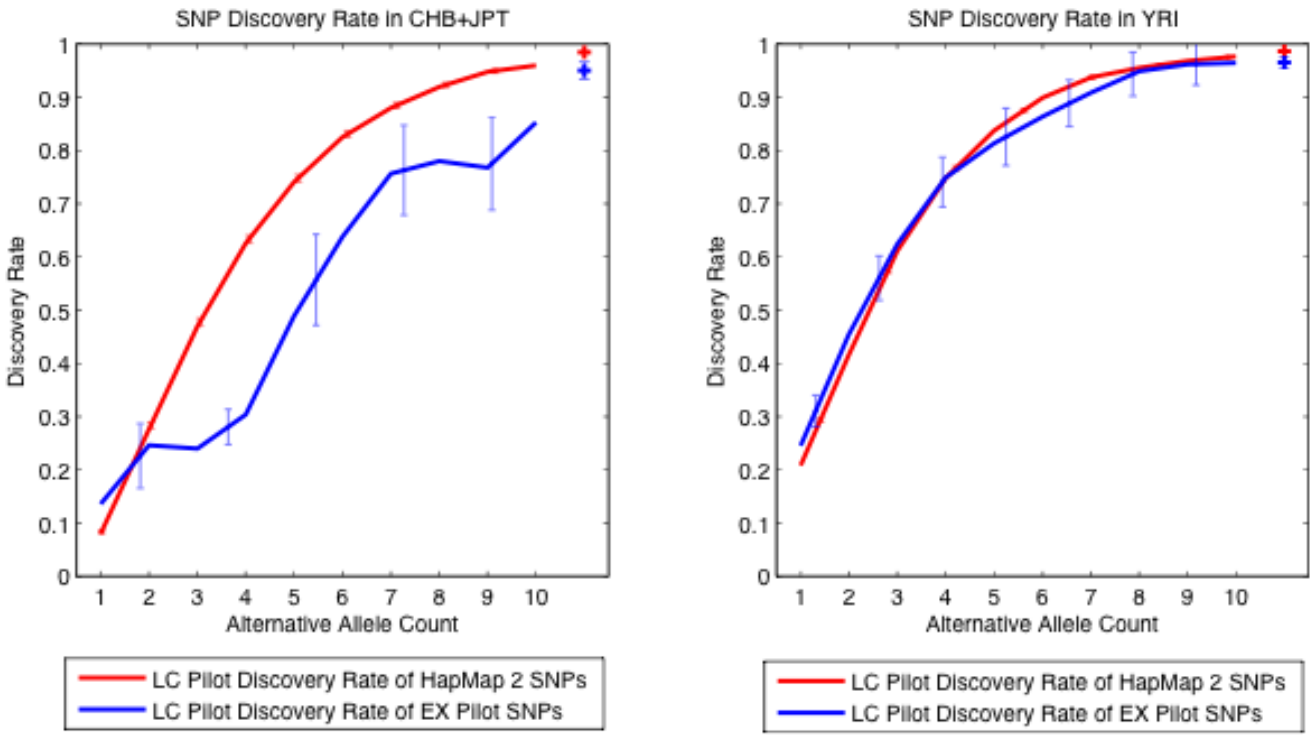
Supplementary Figure 5 – mtDNA haplogroup distribution in the CEU, CHB, JPT and CEU samples. Each continental sample was perfectly divided into population-specific haplogroups, for example haplogroup H for CEU, haplogroup D or B for the East Asian samples, and haplogroup L for YRI.



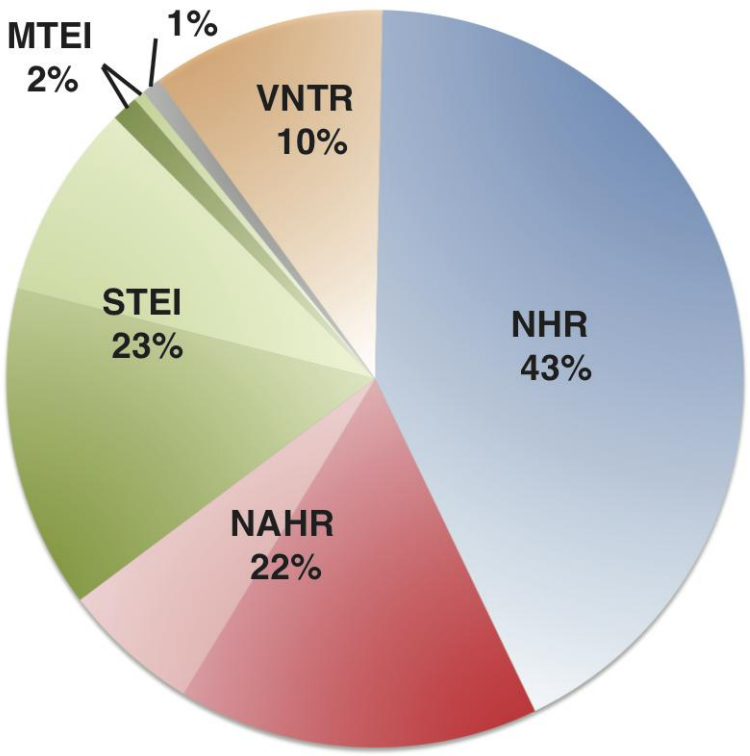
Supplementary Figure 6 - (A) Distribution of heteroplasmy along the length of the mtDNA molecule. (B) Distribution of point heteroplasmy along the length of the mtDNA molecule.



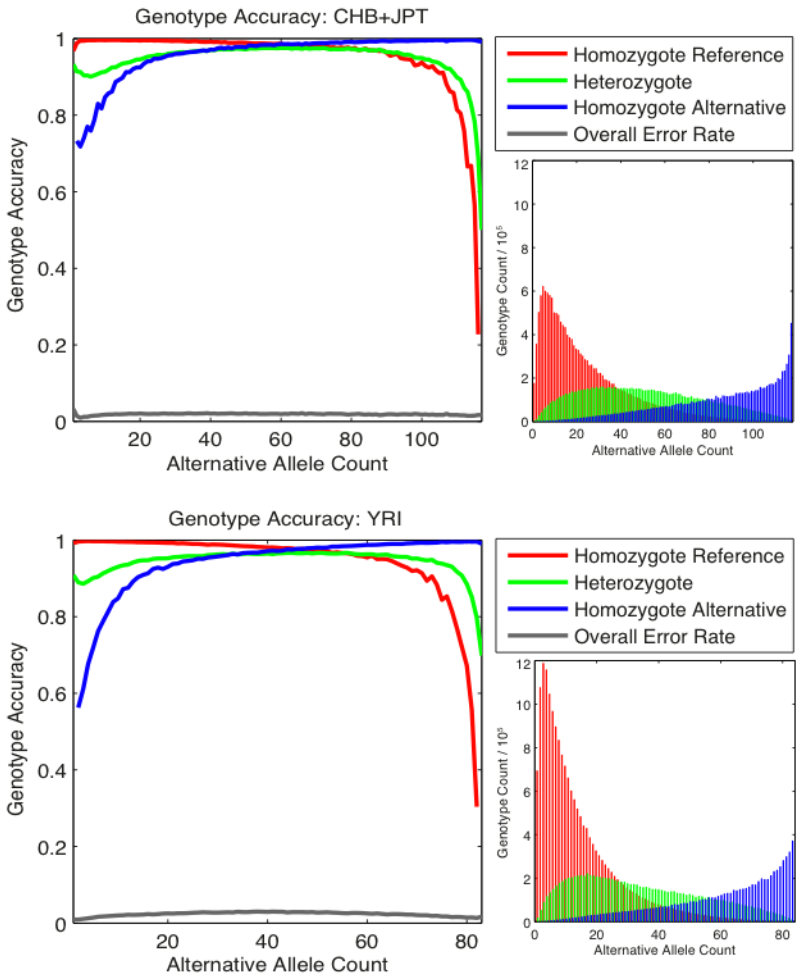
Supplementary Figure 7 - The Y chromosome haplogroup tree, inferred by maximum likelihood from the 2870 variable sites identified. The leaf labels contain the population and the haplogroup assignment of each sample; the latter was obtained from HapMap genotype data.



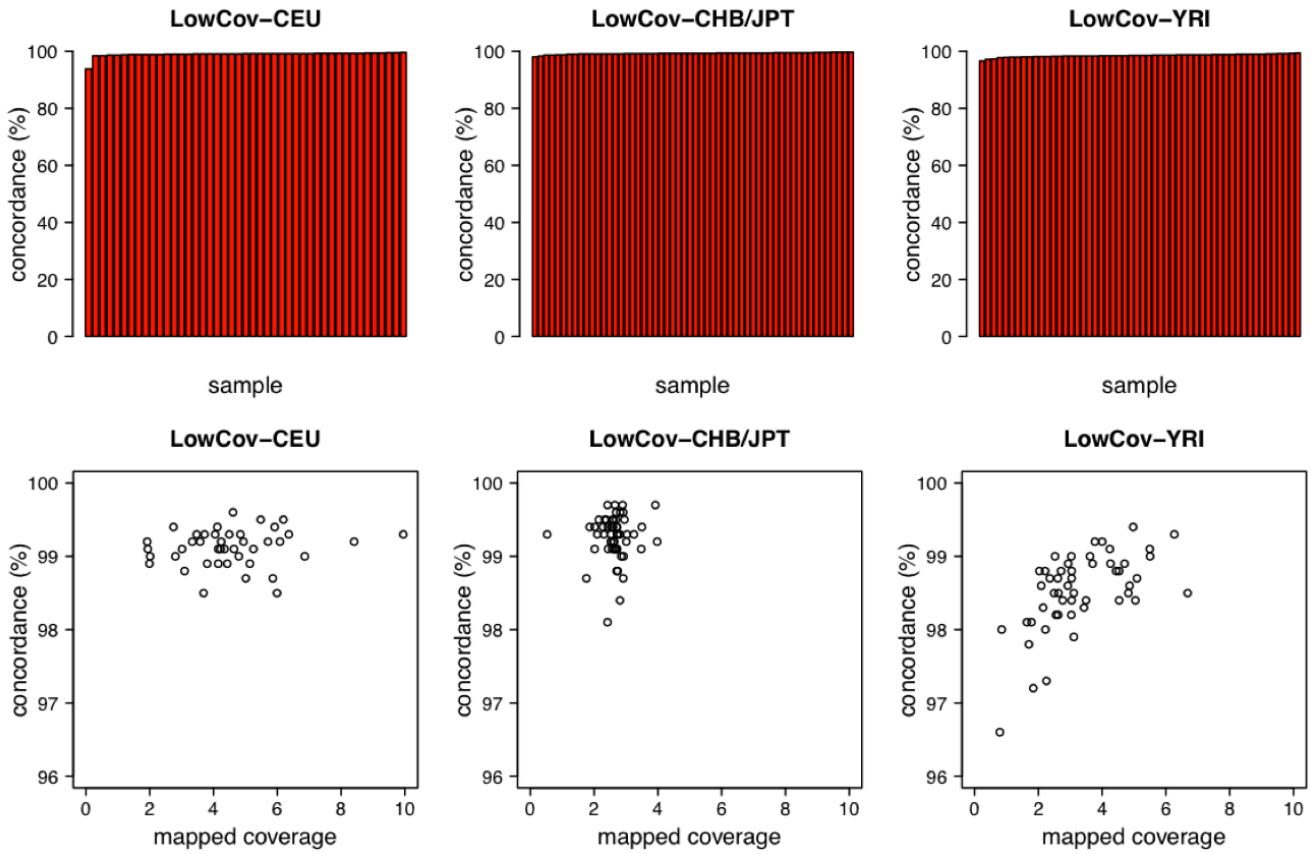
Supplementary Figure 8 - Rate of Pilot SNP Detection Rate by Alternative Allele Count for CHB+JPT (left) and YRI (right). Lines represent the fraction of variants discovered in the low-coverage pilot at a given alternative allele count. Crosses represent the average discovery fraction for all variants having more than 10 copies in the sample. The red lines show the proportion of HapMap 2 sites (excluding sites also in HapMap 3) found to be polymorphic in the Low Coverage Pilot as a function of HapMap alternative allele count. The blue line shows the proportion of Exon Pilot sites found to be polymorphic in the Low Coverage as a function of the Exon Pilot alternative allele count. For both comparisons, only individuals that are present in the intersection are included.



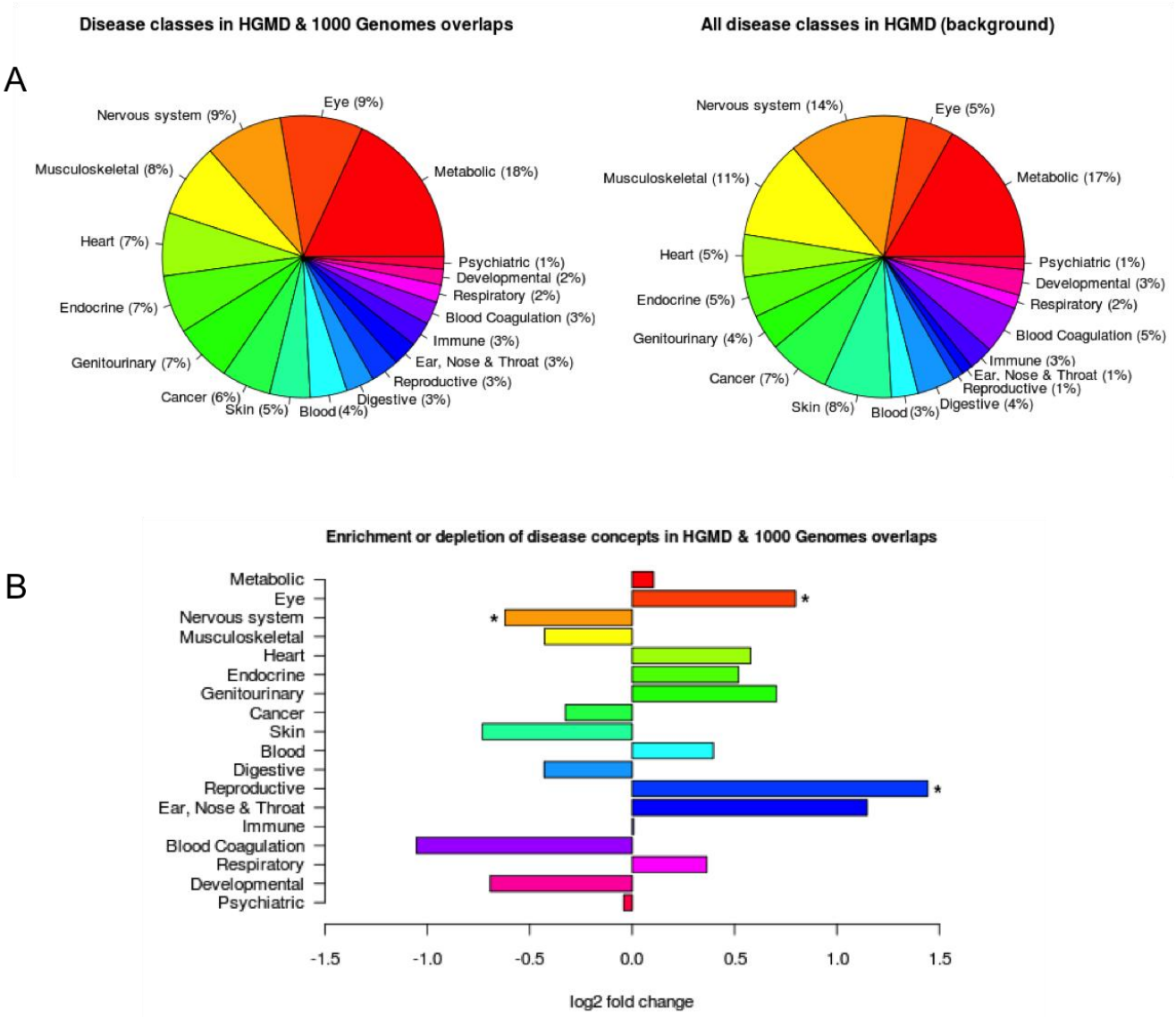
Supplementary Figure 9 - Formation mechanisms of single-nucleotide resolution SVs inferred by BreakSeq. NAHR: non-allelic homologous recombination; VNTR: variable number of tandem repeats; NHR: non-homologous end-joining (NHEJ) or replication fork collapse-associated (FoSTeS/MMBIR); STEI: single transposable element insertions; MTEI: multiple transposable element insertions. In NAHR (red) and MTEI/STEI (green), darker wedges represent high-confidence classification subsets, and lighter wedges are extended subsets.



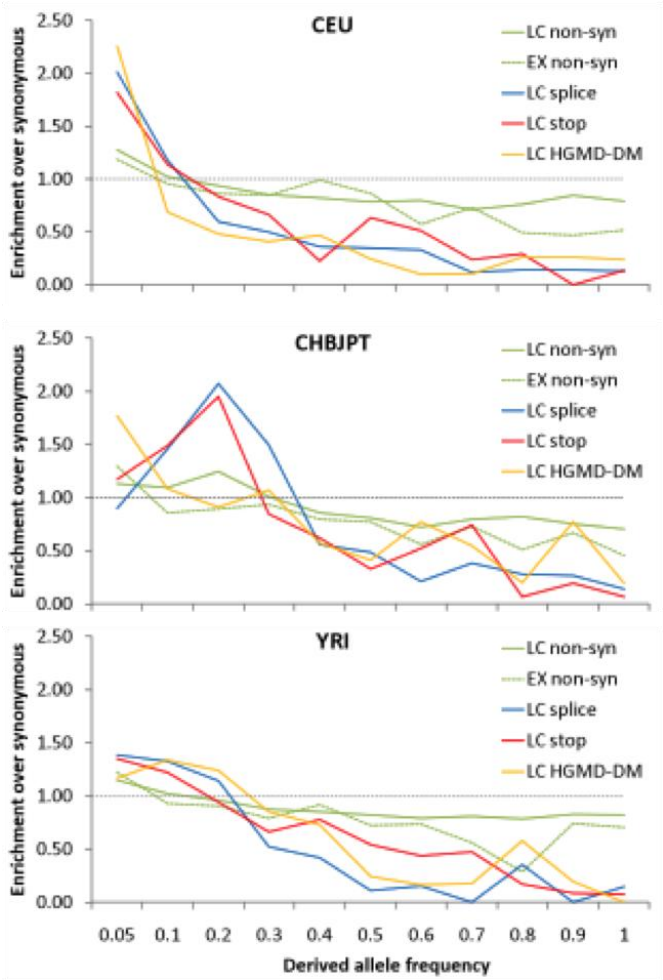
Supplementary Figure 10 - Low Coverage Pilot Genotype accuracy at HapMap 2 sites, not found in HapMap 3, as a function of alternative allele count for the CHB+JPT populations (top), and YRI (bottom). Genotype accuracy is shown separately for homozygote reference calls (red), heterozygote calls (green), and homozygote alternative calls (blue). Also shown is the overall error rate in grey. The number of genotypes in each category as a function of alternative allele frequency is shown to the right of the main plots.



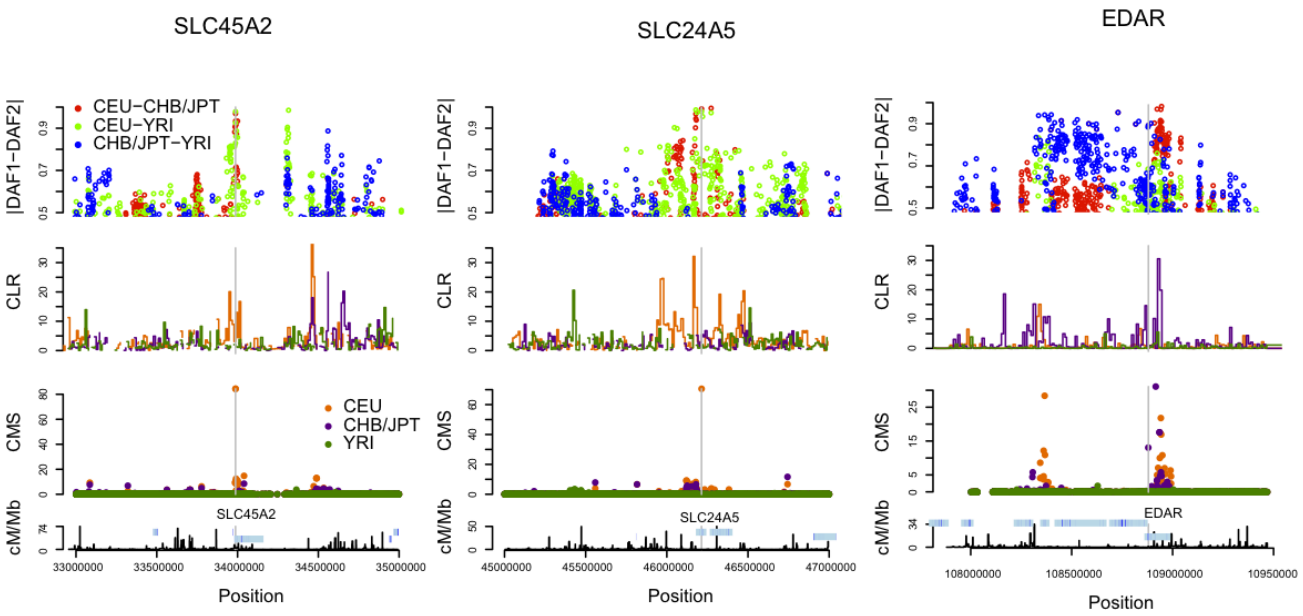
Supplementary Figure 11 - (Top) Bar plots show the concordance of deletion genotype calls with the genotypes of Conrad *et al.* for each low-coverage sample. (Bottom) Deletion genotype concordance plotted versus the mapped coverage for each low-coverage sample by population.



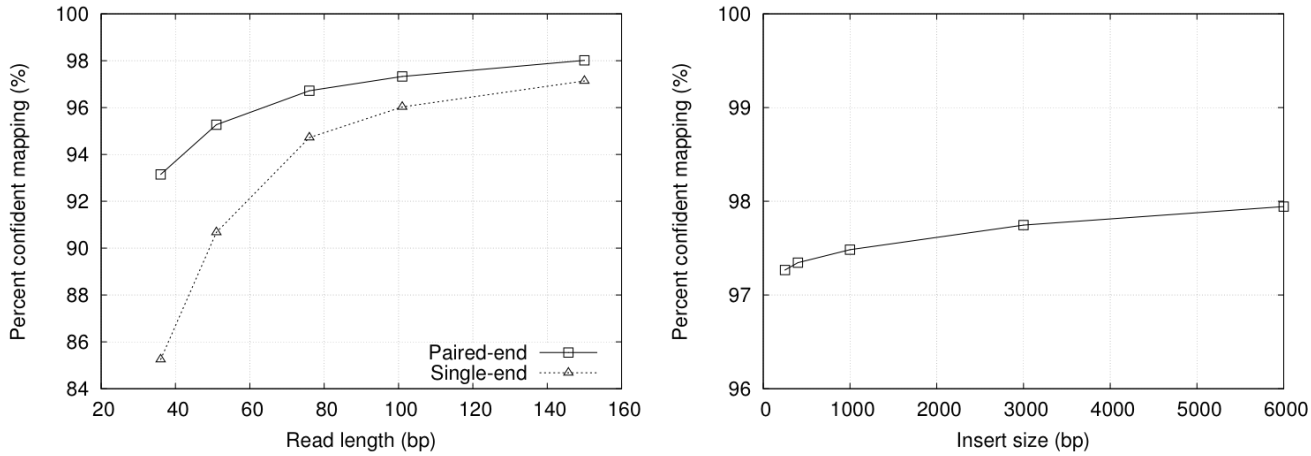
Supplementary Figure 12 - (A) Disease class proportions in the HGMD - 1000 Genomes overlap subset (left) and HGMD background (right), labelled with class and proportion. (B) The ratio of observed to expected HGMD-DM variants found as a function of disease class, where the expected number is based on the distribution between classes in the entire HGMD-DM data set. A star marks classes for which the ratio is significantly different from one ($p < 0.05$ in a Fisher Exact test Bonferroni corrected for 18 tests).



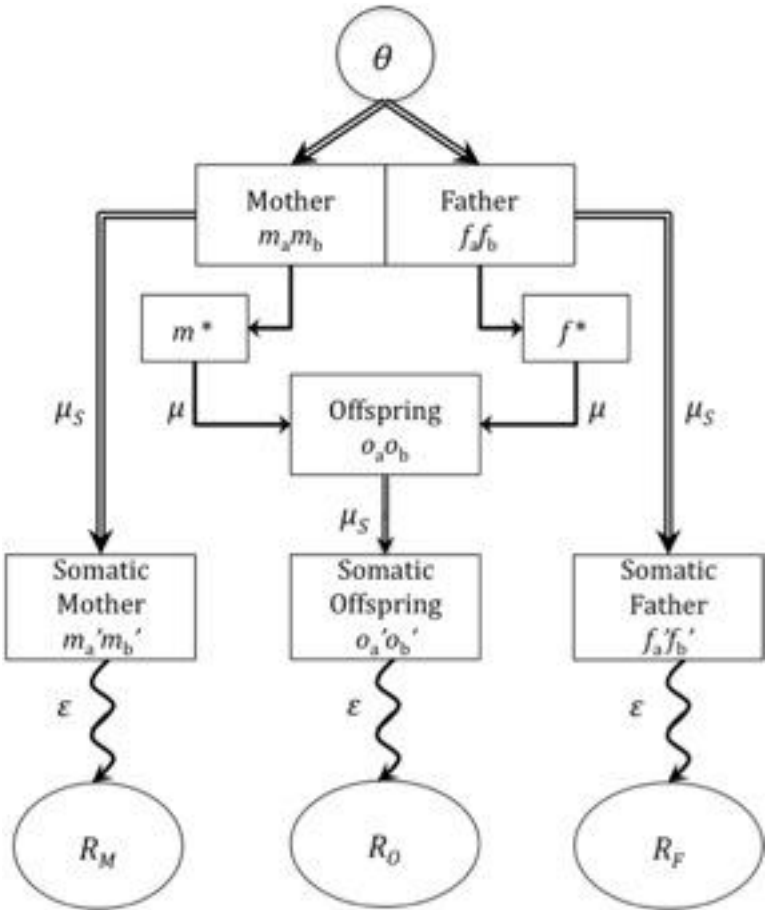
Supplementary Figure 13 - Derived allele frequency spectra for different functional classes of variant, relative to putatively neutral (synonymous) coding variation. Lines indicate the relative proportion of each functional variant class in the specified frequency bin relative to the corresponding proportion for synonymous variants. In general, functional variant classes show enrichment in low frequency bins, although this tendency is most pronounced in CEU. The peak for stop and splice SNPs at a frequency of 0.20 in CHBJPT is likely due to a higher rate of sequencing artifacts in this population, which disproportionately affect functional variation. Weaker enrichment of low-frequency functional variants in YRI is likely due to a combination of lower coverage and weaker LD (leading to poorer ascertainment of low-frequency variants) and poorer HGMD ascertainment of disease mutations in this population. Abbreviations: non-syn, non-synonymous; splice, splice-disrupting SNP; stop, stop codon-introducing SNP; HGMD-DM, variants from the Human Gene Mutation Database classified as "damaging mutations".



Supplementary Figure 14 - Positive selection. Localisation of the signals of local adaptation around three genes previously shown to have strong signals of local adaptation and containing nonsynonymous variants as candidates for the target of selection; the pigmentation genes *SLC45A2(Lamason, Mohideen et al. 2005)* (Phe374Leu at rs16891982) and SLC24A5(The International HapMap Consortium 2005; Sabeti, Varilly et al. 2007) (ALA111THR at rs1426654) and *EDAR(Grossman, Shylakhter et al. 2010)* (VLA370ALA at rs3827760), variants in which are associated with hair and bone morphology(Mou, Thomason et al. 2008; Kimura, Yamaguchi et al. 2009). The plots show, from the top down, SNPs showing strong differentiation in allele frequency between populations, a composite likelihood ratio statistic(Nielsen, Williamson et al. 2005) calculating the evidence for a complete local sweep in each population, the CMS statistic(Grossman, Shylakhter et al. 2010), which aims to localise signals of adaptation, the location of genes and exons (light abd dark blue bars respectively) and the fine-scale recombination rate (HapMap). For both *SLC45A2* and *SLC24A5* the CMS statistic localizes to the nonsynonymous variant, while the population differentiation signal is more diffuse and the CLR statistic peaks away from the variant. In line with previous reports, the strongest signal of selection is around *EDAR* in the CHB and JPT populations. However, two additional features of the signal suggest that the history of selection in the region may be more complex than just a single sweep. First, the signal around the EDAR gene in CHB and JPT is focused 40kb upstream of the coding variant, within the first, untranslated exon and intron and separated by a series of recombination hotspots from the coding variant. Second, there is evidence for two additional weaker selective events: one, as reported earlier(Xue, Zhang et al. 2009), in the same gene in the CEU where the 370A allele is absent, and another, again within the CEU population, focused on the sulfotransferase 1C subfamily gene cluster. Although simulations indicate that a single sweep at the site of the V370A variant can generate high scoring variants 50kb upstream (data not shown), these results suggest a complex history of selection across multiple positions and populations(Coop, Witonsky et al. 2010).



Supplementary Figure 15 - Expected genome coverage as a function of read length with an average insert size of 400bp (left), and insert size with an average read length of 100bp (right).



Supplementary Figure 16 - UdeM trio model for a single site. Sequencing reads covering the site of interest (R_M,R_F,R_O) are the observed data and are indicated by ovals. Neither individual genotypes nor their transmission pattern are observed; rectangles are used to identify these as "missing data". Straight lines are used to indicate allelic lineage; for example, double lines denote that diploid maternal (m_a,m_b) and paternal (f_a,f_b) genotypes are sampled from the population, whereas single lines indicate that each parent contributes a haploid gamete (m^*,f^*) to their offspring. Wavy lines denote where sequencing takes place. Greek letters denote the parameters in the model and have been placed in proximity to the lineages that they affect.

Supplementary Figure References

Coop, G., D. Witonsky, et al. (2010). "Using Environmental Correlations to Identify Loci Underlying Local Adaptation." Genetics.

Grossman, S. R., I. Shylakhter, et al. (2010). "A composite of multiple signals distinguishes causal variants in regions of positive selection." Science **327**(5967): 883-886.

Kimura, R., T. Yamaguchi, et al. (2009). "A common variation in EDAR is a genetic determinant of shovel-shaped incisors." Am J Hum Genet **85**(4): 528-535.

Kin, T. and Y. Ono (2007). "Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat." Bioinformatics **23**(21): 2945-2946.

Lamason, R. L., M. A. Mohideen, et al. (2005). "SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans." Science **310**(5755): 1782-1786.

Mou, C., H. A. Thomason, et al. (2008). "Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form." Hum Mutat **29**(12): 1405-1411.

Nielsen, R., S. Williamson, et al. (2005). "Genomic scans for selective sweeps using SNP data." Genome Res **15**(11): 1566-1575.

Sabeti, P. C., P. Varilly, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.

The International HapMap Consortium (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.

Xue, Y., X. Zhang, et al. (2009). "Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation." Genetics **183**(3): 1065-1077.