SUPPLEMENTARY INFORMATION

doi:10.1038/nature09534

Correction notice

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Nature **467,** 1061–1073 (2010)

In the version of the Supplementary Information originally posted online, section 7.7 was incorrect; an updated version of this section, containing information on calling variants and genotyping exon pilot data is included in the new version of the Supplementary Information. Please see Supplementary Information Table of Contents for details.

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Table of Contents

1. Introduction

In this Supplementary Information we give further information referred to in the main text of the paper. We also provide a technical description of how the 1000 Genomes Project pilot data were collected and processed that will both allow users to understand in more detail how the results were generated and be of use to others analyzing comparable data. In several places to enhance readability we give a summary description first, referring to more complete information later in the document. In many cases, methods and file formats used during the project have been developed by members of the Consortium. References to papers and web-sites describing these resources are provided below.

2. Samples

Requirements for inclusion in the project were: individuals to be sequenced in the project had to have been explicitly consented for broad use and public distribution of extensive genotype or sequence data over the Internet. Cell lines from the samples had to be available to the broad research community, to maximize the value of the 1000 Genomes data by facilitating follow-up studies and use of the sequence data to map cellular phenotypes.

To allow such broad use of complete genome sequence data while minimizing risks to the participants, samples without phenotype data were preferred; samples with phenotype data could be used if those data were available only to close collaborators of the sample collectors and not to the general research community. Samples with pre-existing whole genome SNP and CNV data were preferred for the pilots in order to provide comparative information that would maximize the technical information obtained.

The Samples/ELSI Group and Steering Committee agreed that the HapMap and extended set of HapMap samples (The International HapMap Consortium 2007; The International HapMap 3 Consortium 2010) met these requirements. The HapMap samples came from the Yoruba in Ibadan, Nigeria (YRI); the Centre de'Etude du Polymorphisme Humain (CEPH) collection in Utah, USA, with ancestry from Northern and Western Europe (CEU), the Han Chinese in Beijing, China (CHB), and the Japanese in Tokyo, Japan (JPT), and have proven value as baseline representatives of human genetic diversity. The extended set of HapMap samples had been collected the same way as the HapMap samples (The International HapMap 3 Consortium 2010). They include additional samples from the HapMap populations as well as samples from additional populations. The project includes samples from the Luhya in Webuye, Kenya (LWK), the Toscani in Italia (TSI), and the Chinese in Metropolitan Denver, CO, USA (CHD). The full list of samples for the pilot projects can be found on the project FTP site in the sequence index file. The samples are available to researchers, from the non-profit Coriell Institute for Medical Research (http://ccr.coriell.org/sections/collections/NHGRI/?SsId=11). No identifying or phenotype data are available for these samples.

For the trio project, two HapMap trios were used: YRI daughter NA19240, mother NA19238, and father NA19239; and CEU daughter NA12878, mother NA12892, and father NA12891. Each daughter was chosen, in part, because of extensive prior genomic data including a fosmid library with extensive sequence coverage (Kidd, Cooper et al. 2008) and tiling array CGH data (Conrad, Pinto et al. 2010). The parent samples have HapMap 3 genotype data for 1.6 million SNPs and sequence data in the ENCODE regions (The International HapMap 3 Consortium 2010).

For the low-coverage project, all the suitable unrelated HapMap I and II samples were included, as genotype data are available on 3.5 million SNPs from http://hapmap.ncbi.nlm.nih.gov (The International HapMap Consortium 2007). Some HapMap samples had unexpectedly been shown to be related or have cell line artifacts (McCarroll, Kuruvilla et al. 2008), so were deemed unsuitable. The set was filled out with additional samples from the same populations, many with HapMap 3 data. A total of 179 unrelated samples were sequenced: 60 CEU, 59 YRI, 30 CHB, and 30 JPT.

For the exon project, samples from sets of closely related populations were included to provide as much data as possible on the frequency distribution of low-frequency and rare variants. The trio project samples and most of the low-coverage project samples were included, as well as some from the extended set of HapMap samples, most with HapMap 3 data (The International HapMap 3 Consortium 2010). A total of 697 unrelated samples were sequenced: 105 JPT, 109 CHB, and 107 CHD (321 of East Asian ancestry); 112 YRI and 108 LWK (220 of West African ancestry); and 90 CEU and 66 TSI (156 of European ancestry).

3. Data Generation

Given the early state and diversity of approaches when the project was conceived, the overall work flow for the project was intentionally based upon central sample collection, data generation at multiple sites using multiple technology platforms, iterative development, optimization and comparison of alignment and variant calling routines, and central data submission and final release.

DNA was prepared by Coriell from cell lines and distributed to the nine sequencing centres. All DNA was provided by Coriell in a uniform fashion. We do not know the passage number of the cell lines, but it was not the same for all samples.

The centres performed sequencing using the Illumina Genome Analyzer (I and II), AB SOLiD System (1.0 and 2.0) or the 454 GS FLX. Full details of lanes submitted to the project (including sequencing centre, instrument and library) are given in the sequence index file available on the FTP site. A summary of the amount of data generated by each centre and platform is given in Supplementary Table 1.

In general, data generation followed standard protocols, but the period of data collection for the project (January 2008 through summer 2009) was one of rapid development of the sequencing platforms, so there was some heterogeneity in methods used. Specific details of the methods used in each sequencing centre are given in Section 16 of this document. In brief, sequencing libraries were made by shearing genomic DNA, selecting the required insert size band on a gel, then using the relevant manufacturer's standard methods. Library insert size and concentration were assessed before sequencing using the Agilent Bioanalyzer 2100 and/or quantitative PCR. Sequencing and initial data analysis to extract sequence reads with quality values for each base were performed using manufacturers' protocols and software.

A mix of read lengths and single end (fragment) and paired end libraries were used, as technology advanced and paired end protocols were introduced during the period of data collection. Overall, 77% of the mapped low-coverage data, 80% of the mapped trio data and 56% of the mapped exon data were from read pairs rather than fragments. A more complete breakdown by project and population is given in Supplementary Table 1, and full details of data submission are given in the sequence index file relating to release 2010_07 on the project FTP site.

3.1. Exon Targeting

Exon targeting was by hybridization capture. We intentionally used multiple platforms for capture, as capture reagents and protocols were in an early and rapidly evolving stage of development. All groups received the same initial list of a set of 1,000 CCDS genes. These consisted of 980 randomly chosen CCDS entries and 20 genes that were known to be in ENCODE regions and in parallel analysis in HapMap 3. Due to differences in interpretation of gene models, and the ability of the different capture methods to program design of specific genomic regions, the overlapping targets from the different groups totaled 8140 exons from 906 genes for which coordinates are given on the project ftp site. Details of the approaches used at each of the four centres participating in the exon project are given in Section 16 below, alongside sequence production details. In brief:

- Baylor College of Medicine used NimbleGen 385K capture chips to pull down fragments that were converted to 454 platform sequencing libraries and sequenced single ended on 454 GS FLX/Titanium machines.
- The Broad Institute used RNA baits transcribed in the presence of biotinylated UTP from primers cleaved from an Agilent microarray, capturing the target by pull down with streptavidin beads and sequencing on Illumina GA II machines. Either premade Illumina libraries were pulled down then sequenced, or pulled down material was concatenated then randomly sheared before library making and sequencing.
- The Wellcome Trust Sanger Institute used NimbleGen 385K capture chips to pull down the target region from a premade Illumina paired end library. Captured material was then amplified and sequenced on Illumina GA II machines.
- Washington University in St Louis used a biotinylated capture library generated by PCR in the presence of biotinylated CTP from a pool of 190 bp synthesized oligos to pull down the target region from a premade Illumina sequencing library. Captured material was then amplified and sequenced on Illumina GA II machines.

4. Read Mapping and Generation of BAM Files

Read sets from each sequencing machine run, or lane in the case of Illumina data, were submitted by the sequencing centre to the international short read archives (SRA at NCBI and the ERA at EBI) in SRF format (http://srf.sourceforge.net) with project IDs SRP000031 for the low-coverage project, SRP000032 for the trio project, and SRP000033 for the exon project. From there, the project Data Coordination Centre (DCC, also at NCBI and EBI), picked up the data files, transformed them to FASTQ format (as defined at http://maq.sourceforge.net/fastq.shtml) and made them available on the project FTP site. The DCC carried out basic QC checks on the integrity of the data, and files passing these checks were listed in a date-stamped sequence.index file on the FTP site.
A standard mapping procedure was used for each platform, as described below. Mapped reads were combined into one main BAM file (BAM is a binary representation of the SAM format, http://samtools.sourceforge.net/) per individual per project per platform that contains read alignments, together with a second BAM file of unmapped reads. In the case of read pairs, if one of the pair mapped and the other did not, the unmapped mate was placed adjacent to the mapped mate in the aligned BAM file, as described in the SAM format. The unmapped BAM files were not used during further analysis, but are available from the project FTP site.

The steps taken in the mapping process for Illumina and 454 data were as follows:

- Read sets were downloaded as FASTQ from the DCC.
- Reads were initially mapped to the reference genome.
- Reads from positions of a set of previously-genotyped SNPs were checked to ensure that the sample identifier attached to the sequence was correct. If this check failed the run/lane was removed from the sequence.index file and from all further analysis. Some data on incorrectly labeled samples for which the correct identifier could be deduced were resubmitted to the SRA/ERA, entering the process again at the beginning.
- Quality values were recalibrated, and the adjusted FASTQ files were uploaded back to the DCC FTP site.
- The recalibrated FASTQ files were remapped, using the same parameters as before.
- Lanes from the same library were merged using Picard MergeSamFiles.
- To remove duplicates arising from the library making and cluster calling process (usually only a few percent, but occasionally more abundant), samtools rmdup was run on the lanes produced from paired libraries and samtools rmdupse (single ended) was run on single ended libraries.
- Libraries were then merged to the platform level, and finally Picard MarkDuplicates was run to remove duplicates missed by samtools.

Data generated on the SOLiD platform underwent the same data processing with the following modifications:

- Reads were obtained directly as CSFASTA/QUAL files from the sequencing centres, and not from the Data Coordination Centre (DCC) because the SRF format specifications and convertors were not available for timely submission to the SRA/ERA. SRA/ERA run identifiers were substituted in the BAM files once available.
- Alignments were performed as described below. SOLiD alignments did not include mapping quality scores, were not recalibrated post-alignment, and did not have duplicates removed.

The standard BAM files produced by this process are available from the project FTP site, together with an alignment statistics file containing basic alignment statistics for each BAM file, and an alignment index file containing combined statistics for the entire data set. Mapped data quantities are given in Supplementary Table 1.

These BAM files were used for most subsequent analyses, including all SNP and indel calling in the low-coverage and trio projects, and most structural variation calling. For some structural variation analysis different mapping information was required so different mapping software was used, as described in the relevant subsection of Section 8 below. For SNP and indel calling in the exon project, a second set of mapped BAM files was produced and calls from both sets were combined as described in Section 5.3 below.

4.1. Reference Genome

Reads were aligned to the genome reference NCBI36 including all chromosomes and unlocalised contigs, replacing the mtDNA with the revised Cambridge reference sequence (rCRS; Andrews, Kubacka et al. 1999). Mappings were performed using a sex-specific reference sequence including the Y chromosome for males but not for females. The pseudoautosomal regions of the Y chromosome were masked out. In addition, the Epstein-Barr virus (EBV) genome, used to immortalise the sample cells, was included in the reference.

The reference sequence is available for download from the project FTP sites.

4.2. Mapping of Illumina Data

Lanes were mapped individually using Maq v0.7 with the following parameters: -u (to save all unmapped read pairs to a separate file) and –a 1000 (to mark all read pairs separated by 1000 bp or less as properly paired). All alignments were carried out on a compute cluster consisting of 199 nodes (1192 processors), each with 8-16Gb of RAM.

4.3. Mapping of 454 Data

Lanes were mapped individually using SSAHA v2.4 with the following parameters -454 (which tunes the error model to handle 454 data) and –disk 1 (which saves some intermediate data structures to disk, to reduce memory requirements) on the same compute resource as the Illumina data.

For paired 454 lanes, each end was aligned independently, and the top ten hits for each read were recorded. If both ends aligned uniquely, then the reads were assigned to these positions. If one end mapped uniquely and the other end had multiple hits, then the multiple hit read was placed at the position closest to the expected insert size. If both ends mapped with multiple hits, then the reads were placed at the location closest to the expected insert size and the mapping quality set to zero.

4.4. Mapping SOLiD Data

Alignment of SOLiD data was carried out using the Corona_Lite version v4.0r2.0 pipeline. Sequencing reads were mapped separately to chromosomes 1-22, X, Y, and the mitochondrion allowing up to 3 mismatches. Only mate-pairs mapping uniquely between 500 - 5,000 bp of each other were included within the primary merged BAM files. Local gapped alignment was done using the Small_Indel_Tool for instances where only 1 of 2 mates aligned to the genome. Separate files containing only those rescued gapped mapped reads are available on the project FTP site.

Conversion from Corona_lite version 2 GFF files to SAM format was done using GFF2SAM version 0.01.6. Mapping was done on a cluster of 2688 2.8 Ghz Dell PowerEdge servers with Infiniband DDR.

4.5. Recalibration of Base Quality Values

All data from Illumina and 454 platforms were recalibrated after initial alignment using the following algorithm implemented in GATK software (McKenna, Hanna et al. 2010).

In order to improve the accuracy of the Phred-scaled base quality score Q, we developed a covariate-aware base quality recalibration algorithm that provides empirically accurate base quality scores for each base in every read, adjusted for several significant error covariates such as machine cycle and dinucleotide context. The algorithm first tabulates empirical mismatches to the reference at all sites not known to vary in the population (dbSNP 129), stratifying the bases by their reported quality score (Q_{raw}), their machine cycle in the read, their dinucleotide context, and their read group. For each category we estimated the empirical quality score $Q_{emp} = (N_{mismatches} + 1) / (N_{bases\ observed} + 1)$. These initial $\mathrm{Q_{emp}}$ scores were then broken into linear components and the recalibrated quality score Q_{recal} was estimated according to the equation:

$$Q_{\text{recal}} = Q_{\text{raw}} + \Delta Q_{\text{readgroup}} + \Delta Q_{Q_{\text{raw}}} + \sum_{\text{covariates}} \Delta\Delta Q_{\text{covariate},Q_{\text{raw}}}$$

where each ΔQ and $\Delta\Delta Q$ are the residual differences between empirical mismatch rates and those implied by the reported quality score for all observations conditioning only on the Q_{raw} or on both the covariate and the Q_{raw}, as indicated. A second pass through the reads updated the quality scores to their Q_{recal} values given their Q_{raw} scores and covariate context, enabling downstream tools to benefit without modification.

To quantify the impact of recalibration, we called SNPs for NA12878 (the CEU trio daughter) with both raw and recalibrated data using the GATK SNP Caller. In the recalibrated data, the total number of variants called decreased by 2.8%, with the net calls removed having a transition to transversion (Ti/Tv) ratio of 1.07, compared to 1.96 for the post-calibration calls. Typically, true variants have a Ti/Tv ratio around 2, whereas random changes in a genome with all four bases equally frequent have Ti/Tv of 0.5.

4.6. Comparison of Read Data to Known HapMap Genotypes

In order to check that the sequenced data had been correctly assigned to each individual, the sequenced genotypes were matched to the genotypes from HapMap II for the same samples (The International HapMap Consortium 2007). After each sequence lane was mapped to the reference, genotype log likelihoods were calculated with the samtools pileup –g command, and using these the log likelihood of the data for each set of HapMap 3 genotypes was calculated. Sequence runs/lanes for which either the best matching genotypes were not from the expected individual, or the best matching genotypes did not separate well from all the other individuals (factor of 1.2 separation) were rejected and removed from all subsequent analysis.

Across all three projects 650 of 13,042 SRA run ids originally submitted to the project were rejected because of bad likelihoods (5%). In addition 126 run ids were dropped from analysis because of other quality control failures, in most cases extremely low data quality or missing data, and 121 were withdrawn for other reasons (2% total), leaving 12,145 (93%) contributing to the project analysis.

5. SNP Calling

Whereas a single read mapping process was used for most of the analysis in the project, multiple SNP calling procedures were used. In part this was because different methods were thought most appropriate for the design of three different data sets, in part because different approaches were under active development by individual members of the project during the course of the pilots, and in part because we found empirically that, given the state of current methods, the consensus of multiple primary call sets from different methods proved to be of higher quality than any of the primary call sets themselves (Section 5.1.5).

All SNP callers used in the project share several features. Each caller starts by examining the sequence reads overlapping each site in the genome. Then, base calls and quality scores overlapping each position are examined and used to calculate the probability of observed bases and quality scores for each individual given a potential underlying genotype. We call these quantities the “genotype likelihoods” defined as:

$$\mathrm{GL}_{ij}(g) = \mathrm{P}(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | \mathrm{G}_{ij} = g)$$

Here, GL_{ij} is the genotype likelihood associated with a specific genotype g and position j in individual i. It is calculated as the probability of observing the vectors of bases $\mathbf{B}_{ij}$ and $\mathbf{Q}_{ij}$ overlapping position j in the mapped reads for individual i. Typically, we considered 10 values for the underlying genotype g, corresponding to the 10 possible genotypes (A/A, A/C, A/G, A/T, C/C, C/G, C/T, G/G, G/T and T/T). Standard equations to calculate the GL_{ij} values are given in Li and Durbin (2008), together with a modified version that allows for simple dependency in the relationship between base qualities of multiple reads at a site; the equations given there are implemented in samtools, with the default being to use the dependency model with theta = 0.85. The resulting likelihoods are used by several of the SNP callers described below.

To assign genotypes, all methods then use Bayes rule to assign individual genotypes as:

$$\mathrm{P}(\mathrm{G}_{ij} = g | \mathbf{B}_{ij}, \mathbf{Q}_{ij}) = \mathrm{P}(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | \mathrm{G}_{ij} = g)\, \mathrm{P}(\mathrm{G}_{ij} = g) / \mathrm{K}_{ij}$$

Where K is a normalizing constant, defined as $\mathrm{K}_{ij} = \Sigma_g \mathrm{P}(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | \mathrm{G}_{ij} = g)\, \mathrm{P}(\mathrm{G}_{ij} = g)$. Although there are some differences in the calculation of the likelihoods $\mathrm{P}(\mathbf{B}_{ij}, \mathbf{Q}_{ij} | \mathrm{G}_{ij} = g)$, the main way methods differ is in the types of information and the approaches used to estimate $\mathrm{P}(\mathrm{G}_{ij} = g)$, the prior probability for genotype g at each location j in individual i.

Once an initial set of variants is called, a number of post-processing steps are taken to remove potential false positives. All sets of variants are stored in VCF format (http://1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcf4.0), which as well as representing the location on the genome and nature of the base change, also supports an identifier, a quality score scaled in PHRED ($-10\log_{10}$) units analogous to a base quality, a possible filter field, an extendable field containing additional information in tag=value notation, and the genotypes of the sequenced individuals with their own qualities and ancillary information.

Filters that were applied include:

- Depth thresholds to reject SNPs where the read depth in at a site is either much lower or much higher than expected, suggestive of copy number variation that might lead to miscalling of SNPs based on the mapping of paralogous sequences. We rejected sites that were more than twice or less than half the mean depth across samples in a population/trio in the low-coverage and trio genome-wide data sets (depth is too variable in exon data to use this filter).
- Local realignment to remove false positives due to small insertions and deletions (indels). Current mapping algorithms are limited in that they align each read independently, without considering all overlapping reads jointly to guide placement of variants. Consequently, reads whose first or last few bases overlap an indel tend to have those bases misaligned as mismatching bases rather than indels (because the gap open penalty is larger than the mismatch penalty in typical mapping algorithms). In order to correct apparent SNP calls in fact caused by misalignment around an indel, we applied a local realignment process as a filtering step around each candidate SNP call in the genome-wide data sets.
- Poor mapping quality. Because the reference genome is both incomplete and does not represent all genetic variation, reads from unrepresented regions can end up mapping with apparent certainty to incorrect locations, thus leading to false positive variant calls. Although this remains a challenge to variant calling, such regions are often highly repetitive. Our experience led us to identify regions where a substantial fraction of reads map with low mapping quality as being strongly enriched for false positive calls (Section 6.1). On this basis we constructed a map of the accessible genome, details of which are given below in Section 5.1.6 and discussed in the main text.

Below we first describe the specific methods used to call SNPs in low-coverage data, then the differences in calling variants in the deeply sequenced trios, then the exon data (all exon targets were autosomal).

5.1. Low-Coverage SNP Calling

Calling was separate for the three analysis panels, CEU, YRI and CHB+JPT. For each, three primary SNP call sets were generated, from the Broad Institute (Broad), the University of Michigan (Michigan) and the Sanger Institute (Sanger). Details of the methods are given in separate publications (DePristo, Banks et al. 2010; Le and Durbin 2010; Li, Willer et al. 2010), with brief details of the main features given below. All three were produced by a two step processes, involving in the first step a set of candidate calls made based on the evidence at each base pair in the reference, independent of neighbouring sites. Then in each case a more computationally intensive linkage disequilibrium (LD)/imputation based approach was used to refine the call set and genotypes for all individuals. Formally, the difference between the first and second steps can be seen as a difference in the prior P(G_{ij} = g); the likelihoods remain the same.

5.1.1. Generation of Broad Low-Coverage SNP Call Set

The GenomeAnalysisToolkit (GATK, http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) module UnifiedGenotyper (DePristo et al 2010; DePristo, Banks et al. 2010; McKenna, Hanna et al. 2010; McKenna et al 2010) was used to calculate genotype likelihoods from all samples and technology platforms in an analysis panel simultaneously. This calculation was the same as that used in samtools assuming read independence, but only used reads with a minimum mapping quality at least 10 and fewer than 4 mismatches within 40 bp, mate pairs mapped to the same chromosome, and bases with quality score greater than 10.

SNP candidate sites were called assuming an unstructured population of unrelated individuals with a per site prior probability of polymorphism set at 0.001. An E-M algorithm is used to estimate the allele frequency at each site by maximum likelihood and candidate sites require a posterior probability greater than 0.9 (corresponding to a minimum PHRED scaled quality score of 10).

In order to detect technical artifacts caused by systematic sequencer errors, strand-specific log odds values (LODs) were computed using only non-reference evidence from the forward strand, and separately for the reverse strand, divided by the all-reference hypothesis. If either of these strand-specific LODs, referred to as SLOD, is positive then the non-reference alleles have a directional bias. Frequency dependent filter thresholds for LOD and SLOD were fit to maximize the number of candidate SNPs passing a minimum transition to transversion ratio threshold.

Finally, LD/imputation based genotype calling was carried out at each candidate site using the BEAGLE package, which can take genotype likelihoods as described above and return an estimate of the true genotypes, as described in Browning and Yu (2009).

The Broad call set contained SNP calls and genotypes for the X chromosome as well as the autosomes, but the X chromosome calls used the same methods as the autosomes and were not sex aware, so potentially modeled heterozygous sites in males.

.

5.1.2. Generation of Michigan Low-Coverage SNP Call Set

For the Michigan call set, genotype likelihoods were calculated on a per platform basis using the default samtools genotype likelihood model, as described above. For each platform, sites where alignment depth was too high (top percentile) or too low (bottom percentile) were excluded from analysis. The resulting platform-specific genotype likelihoods were then combined for each individual. This strategy effectively assumes dependency between base calling errors within a platform, but no dependency across platforms.

To identify candidate polymorphic sites, we used Brent's likelihood optimization algorithm to estimate allele frequencies at each locus. When comparing $L_{\text{noVariant}}$ to L_{variant} we favored sites with transition polymorphisms over those with transversions (in our prior, 2/3 of polymorphic sites are expected to be transitions). Sites were considered as potentially polymorphic when the posterior probability of a variant call was ~0.90 (corresponding to a phred scaled quality score of 10).

For the LD based refinement step the Michigan SNP calls used the Markov model implemented in MACH (http://www.sph.umich.edu/csg/abecasis/MACH) that decomposes the haplotypes carried by each individual into a series of segments, each copied from a different individual. The model tries to minimize the number of switches between segments and the number of discrepancies between the genotypes implied by each segment and observed read and genotype data for each individual. The model is described in detail elsewhere (Li and Stephens 2003; Li, Willer et al. 2010).

The model proceeds stochastically: initial estimates of the haplotypes for each individual are generated, the genome of each individual is decomposed into haplotype segments derived from the other individuals, and these haplotype segments are used to derive a prior for each genotype at each position. After updating genotype call estimates, the procedure was repeated 100 times. Each iteration of this procedure generates a pair of estimated haplotypes and genotypes per individual, which were then used to generate a set of consensus haplotypes and genotypes (Li, Willer et al. 2010).

A final, method-specific filter was applied. For each marker, we calculated an r^2 statistic that estimates the correlation between estimated genotype calls and the true underlying genotypes (Li, Willer et al. 2010). The r^2 measure is defined as the ratio of the variance in expected genotype counts for the minor allele (a real number between 0.0 and 2.0 for each individual) and *2p(1-p)*, the expected variance of this quantity in a Hardy-Weinberg equilibrium population where genotypes are observed without error (here *p* is the estimated allele frequency for the site of interest). This measure has been used to filter poorly imputed markers in genome-wide association scan meta-analysis (Demichelis, Setlur et al. 2009). We filtered out sites with r^2 estimates of 0.5 or less.

The Michigan low-coverage calls were for autosomes only.

5.1.3. Generation of Sanger Low-Coverage SNP Call Set

The Sanger low-coverage genotype calls started from the samtools likelihoods.

The subsequent candidate generating step and LD/imputation based post analysis are implemented in the software package QCALL (Le and Durbin 2010). Candidate polymorphic sites are found by a dynamic programming algorithm that estimates the probability of the data given there are k non-reference alleles in 2N chromosomes (for all k). For each site, the probability of a SNP is calculated as the probability of k > 0, assuming a prior probability for the variant frequency of $p(k) = \theta(1/k+1/(2N-k))$, where θ is the per-site population mutation rate (taken as 0.001 for humans). This generated approximately 40 million SNP candidates across all 179 samples combined.

In the second, linkage disequilibrium aware, analysis, shared haplotype structures are used to estimate posterior probabilities of SNPs and genotypes. For each population, CEU, YRI, and CHB+JPT, 20 possible ancestral recombination graphs for the full set of samples were built using MARGARITA (Minichiello and Durbin 2006) on genotypes or, where available, phased haplotypes from the HapMap project (The International HapMap 3 Consortium 2010), considering 1 Mb of the genome at a time. For each SNP candidate, 40 marginal ancestral trees inferred at the left and right flanking genotyped sites were used to estimate the SNP posterior probability by evaluating the likelihood of the observed sequencing data for all possible (single) mutations in the 40 trees. Genotypes and phased haplotypes are estimated by integrating over the different trees and sites within each tree. SNPs were subsequently filtered by removing regions containing three or more calls within 10 bp (FW10 filter).

The Sanger low-coverage calls were for the autosomes and the X chromosome and were sex aware so that the male X chromosome was treated as haploid.

5.1.4. HapMap 3 Genotype Data as Scaffold for Analysis of Low-Coverage Data

Nearly all samples selected for sequencing by the project had previously been genotyped by the HapMap 3 project. We used HapMap 3 genotype data in our analysis in three ways.
First, by comparing reads in each lane to HapMap 3 genotypes that they putatively overlap, we used the HapMap 3 genotypes to very sample identities and minimize mistakes in sample tracking (Section 4.6).

Second, for the Michigan low-coverage call sets, we used HapMap 3 genotypes to aid analysis of low-coverage data. Specifically, we adjusted the genotype likelihoods at HapMap 3 sites to consider both the observed read data at that location and the observed HapMap 3 genotype. Thus, at HapMap 3 sites our likelihood become $GL_{ij}(g)$ α P(HapMap 3 Genotype|$G_{ij} = g$) * P(Base Calls, Quality Scores|$G_{ij} = g$). The definition of P(HapMap 3 Genotype|$G_{ij} = g$) assumed a small error rate (of about 0.2%) at HapMap 3 sites. Supplementary Table 13 shows that, in simulated datasets, using a scaffold of genotyped SNPs in this way is expected to improve genotype call accuracy slightly, particularly when sequencing depth is low.

Third, for the Sanger call set, haplotypes derived from analysis of HapMap 3 were used to construct a model of haplotype variation across each sequenced region on which low-coverage genotype calls were then placed. In this way, for example, if a set of individuals was predicted to share a particular haplotype segment identical-by-descent based on HapMap 3 information, the Sanger caller favored solutions that assigned a consistent allele to all these individuals. Since HapMap 3 haplotypes were for the CEU and YRI samples were derived using trio information, they are expected to be extremely accurate. In effect, for CEU and YRI, this analysis places genotypes derived from analysis of sequence data onto a framework of extremely accurate haplotypes derived from trios. As with the Michigan call set, Sanger calls are expected to be much more accurate at HapMap 3 sites than elsewhere in the genome.

An important consequence of the last two uses of HapMap 3 genotype data is that genotype calls derived by the project at HapMap 3 sites are generally more accurate than those generated elsewhere in the genome. Thus, in all evaluations of genotype call quality (for example, in comparisons to HapMap II genotype data), we specifically focus on sites not included in HapMap 3. Outside of HapMap 3 sites, similar to the simulations presented in Supplementary Table 13, we expect a much smaller improvement in accuracy results from using the HapMap 3-based model of haplotype variation.

5.1.5. Merging of Low-Coverage SNP Call Sets

Analysis of individual call sets produced by Broad, Michigan and Sanger demonstrated that a consensus approach to defining variant positions led to a higher quality data set as demonstrated by: (a) genotype calls that were more accurate than in any single call set, (b) a transition-transversion ratio at newly discovered SNPs that was more similar to that observed at dbSNP sites, (c) an overall transition-transversion ratio that was closer to value of slightly >2 observed in previous SNP discovery efforts, (d) consistently high rediscovery rates for dbSNPs and HapMap SNPs (Supplementary Table 12). Overall, genotypes obtained through a consensus procedure are estimated to have 30% fewer errors than those generated by any single caller (Supplementary Figure 1).

Consensus genotypes were defined using a simple majority vote among callers. In case of conflicting evidence (for example if a different genotype was listed in all three sets or the genotype was different in two sets and absent from the third), the following order of precedence was used: Sanger, UMich, Broad for the CEU and YRI call sets, and UMich, Sanger, Broad for the CHB+JPT call set. Where possible, phase was taken from the primary dataset. If the chosen genotype was not from the primary dataset, then the genotype was set as unphased. SNP and genotype quality scores were not retained in the merged dataset.

Low-coverage SNP calls on the X chromosome were derived from call sets generated by the Broad and Sanger pipelines described above, and only calls made by both pipelines were reported. At sites in the intersection, the genotypes reported came from the Sanger calls, which were sex aware.

The primary call files as well as the final merged files are available from the project FTP site.

5.1.6. The Accessible Genome

Following merging, additional filters were imposed on the set of SNPs on a per-population basis. These were (1) the average (per sample) coverage at a candidate SNP had to be within a factor of 2 of the median genome coverage, and (2) the fraction of reads mapping to the candidate SNP location that have a mapping quality of zero had to be less than 10%. The implications for the fraction of genome that is accessible using these metrics are discussed in the main text.

The filters on coverage and fraction of reads with low mapping quality described above lead to the exclusion of a substantial fraction of sites in the genome. To enable population genetic analysis and to describe the completeness of the resource we created mask files that define, for each population/analysis panel, those bases that were accessible to SNP discovery in the low-coverage project (available on the project FTP site). Overall, about 15% of the genome was excluded through the filters, with the fraction varying slightly between analysis panels (a result of different coverage and technology combinations). Details of the accessible genome for CEU are given in Supplementary Table 2.

5.1.7. Phasing of Merged Low-Coverage SNP Call Sets

The call set merging produced a set of partially phased haplotypes, i.e., a series of haplotype fragments that were phased relative to each other but separated by unphased heterozygotes. We sought to “phase-finish” these haplotypes by using an LD-based method to place the remaining unphased alleles onto the haplotype scaffold that resulted from the merging.

To do so, we used a modified version (v2.2) of the IMPUTE2 software (Howie, Donnelly et al. 2009). IMPUTE2 has previously been described in the context of genotype imputation from a reference panel, but the algorithm also includes a phasing step, and this can be used to produce best-guess haplotypes from unphased genotype data. Howie et al. (2009) explain the basic procedure for sampling from the joint posterior distribution of haplotypes underlying the genotypes of a number of individuals: the IMPUTE model (Marchini, Howie et al. 2007) is embedded in a Gibbs sampler, and at each iteration every individual samples a new pair of haplotypes, conditional on the current guesses of the other individuals.

We used a similar procedure here. The main difference is that we assigned different HMM emission probabilities to the phased and unphased genotypes: unphased genotypes used the emission probabilities described by Marchini et al. (2007) and Howie et al. (2009), whereas phased genotypes used simpler probabilities that result from ordering the observed alleles. Supplementary Table 1 of Marchini et al. (2007) shows the probabilities for “mutating” from two ordered copied alleles to two unordered observed alleles; when the observed alleles are ordered (i.e., phased), only the column labeled ‘1’ (observed heterozygote) changes.

In addition to sampling from the posterior distribution of haplotypes, we also calculated the marginal probabilities of both possible phase calls at each unphased heterozygote. When summed across iterations, these Rao-Blackwellized probabilities amount to weighted Monte Carlo counts of the phase calls. At the end of a run, we divided these counts by the number of iterations (minus burn-in) to generate marginal posterior probabilities. In order to generate best-guess, phase-finished haplotypes, we simply chose the phase call with the largest marginal probability at each unphased genotype. In the rare case that the posterior probabilities were both exactly 0.5, we phased that genotype at random.

We omitted singleton SNPs from this analysis because they contain very little LD information for phasing; we also omitted SNPs with more than two alleles since they are not easily handled by the model. We phased each analysis panel separately, and we analyzed each chromosome in non-overlapping 5 Mb chunks for computational convenience. Each of these regions had a 250 kb buffer (which was used for inference but omitted from the output, and which did overlap between chunks) to prevent edge effects. We combined the chunk-specific output files into whole-chromosome haplotypes by simple concatenation – since the phase-finishing was applied to a chromosome-wide haplotype scaffold, there was no need for more complicated ligation procedures. For each phasing analysis, we ran IMPUTE2 for 110 iterations, the first 10 of which were discarded as burn in. Since this is a small sample of individuals, we did not use the state selection approximation that is usually used with IMPUTE2; i.e., we set the number of HMM states (-k in the software) to be the entire sample.

Although monomorphic sites, singleton sites and tri-allelic sites were excluded from the phase-finishing, they were replaced into the phased haplotypes available from the project FTP site so that the files agree on a line-by-line basis with the partially-phased genotype files.

5.2. Trio Data SNP Calling

Trio SNP calling was analogous to low-coverage calling except that only two call sets were generated, from Broad and Michigan, and the LD/imputation based refinement stage that was relevant to population data was simply removed in the Broad set, which relied on much deeper data, and replaced in the Michigan set by a structured prior that enforced Mendelian segregation. The intersection of the two call sets was taken as the consensus. Details are given below.

5.2.1. Generation of Broad Trio Call Set

A pre-processing step used GATK to locally realign all sequence reads that cover either an indel or a cluster of apparent polymorphisms (DePristo, Banks et al. 2010; McKenna, Hanna et al. 2010). Likelihoods were calculated as for the low-coverage data, and the likelihood-based SNP calling treated each trio as a population of three unrelated females.

Because coverage is high a minimum threshold of Q50 was required before declaring a potential variant site. These raw SNP sites were subsequently filtered based on the following criteria:

- Heterozygote Allele balance (AB) <= 75% reference
- Depth of coverage (DP) <= 360, AND
- Strand bias (SB) <= -0.10, AND
- (Number of covering reads with mapping quality score zero (MQ0) <= 0.1 * depth of coverage OR (MQ0 < 4)).

Only SNPs passing all four of the above criteria were included in the final Broad Institute (BI) call set. Full descriptions of these filters can be found in the documentation for GATK.

5.2.2. Generation of Michigan Trio Call Set

Likelihoods were obtained from SAMtools, and combined using a prior that enforced Mendelian segregation. Genotypes were phased where possible using transmission among family members, otherwise reported as unphased. A potential polymorphic site had to satisfy:

- For each sequencing technology, the depth of coverage combined across all trio members at the site is between 50% and 150% of the average depth, and the root mean squared (RMS) mapping quality score of covering reads is at least 30.
- Sequence data passing the above filter is present for each person in the trio (although perhaps not from the same sequencing technology).
- The posterior probability for at least one non-reference allele exceeds 0.999 (SNP call quality score > 30).
- The site is at least 3 bp away from a potential indel detected by local realignment.

The primary call set integrates data from all sequencing platforms. Additional call sets, specific to each platform, were also generated to enable analysis of the extent of overlap in technologies.

5.2.3. Merging of Trio Call Sets

The release set consists of all sites that are called by both analyses, including passing the criteria shown above. In almost call cases where both approaches called the same SNP, genotypes for the three family members were identical. Where there were differences, genotypes from the trio-aware Michigan call set were used. As with the low-coverage analysis a fraction of the genome had to be excluded due to high or low coverage and poor mapping quality. Masks for each trio are available on the FTP site and a summary of the accessible genome is given in Table 1. Because the trio project data were typically generated earlier than the low-coverage data, a greater fraction of the genome has to be excluded (approximately 20%).

5.2.4. Phasing of Trios

Once the trio genotypes were called, we set aside all SNPs that violated the rules of Mendelian inheritance (these were later used in the analysis of *de novo* mutation and structural variation). We then phased as many of the remaining SNPs as possible by identifying unambiguous transmissions from at least one parent. SNPs that are heterozygous in the child and both parents cannot be phased in this way, and we sought to recover some of these by borrowing LD information from the low-coverage project.

For each trio, we used IMPUTE2 to phase the intersection of the SNPs called in the low-coverage project dataset and the corresponding trio project dataset (CEU and YRI, respectively). We fixed the phase of the low-coverage haplotypes (following the phase-finishing step described earlier), as well as the haplotypes of the trio parents at SNPs with unambiguous phase. We then phased the remaining SNPs in the parents, separately for CEU and YRI, as described in section 5.1.7 on phase-finishing the low-coverage project.

The trio parents were phased as if they were unrelated, in the sense that the model did not force the parents to transmit opposite alleles at SNPs that were heterozygous in both parents and their child. Instead, this constraint was enforced post-hoc by recalculating the marginal posterior probabilities for each unphased genotype, conditional on the parents transmitting opposite alleles.

5.3. Exon Project SNP Calls

The exon project SNP call set was composed of the intersection of SNP calls made using the GATK Unified Genotyper essentially as described above for low-coverage and trio calls, with calls from an alternate pipeline using the MOSAIK read mapper and the GigaBayes SNP caller. Note that unlike the low-coverage and trio projects, the different SNP call sets started with different read alignments. SNP calls were made separately for each of the 7 exon project populations.

5.3.1. MOSAIK Read Mapping for the Exon Project

Reads for Illumina and 454 data were mapped using the MOSAIK read mapper (http://code.google.com/p/mosaik-aligner/). This uses gapped alignment to map reads and is expected to have higher sensitivity for reads containing indels than MAQ. The hash size used was 15 for Illumina reads with a minimum mismatch of 4 (36mer read length), 6

(51mer read length), and 12 (76 and 101mer read lengths). Parameters for alignment of 454 reads consisted of a hash size of 15 with at least 70 percent of the read being aligned and the maximum percentage of mismatches of 5 %. Following alignment, GATK was used for quality score recalibration as described above. Duplicates were removed using Picard MarkDuplicates for Illumina reads and BCMMarkDuplicates for 454 reads.

5.3.2. Broad Institute SNP Calls

The Broad SNP calls were made separately within each of the seven ESP populations. A superset of calls was obtained as the union of all per-population calls.

An initial step used GATK to locally realign all sequence reads that cover either an indel or a cluster of apparent polymorphisms (DePristo, Banks et al. 2010; McKenna, Hanna et al. 2010). The likelihood based SNP calling treated each trio as three unrelated females. The GenomeAnalysisToolkit (GATK) module UnifiedGenotyper was used to calculate genotype likelihoods from all samples and technologies simultaneously, exactly as in the low-coverage analysis, except that potential SNP sites were only retained when the posterior probability of a segregating SNP exceeds a minimum threshold of 10^{-3} (Q30).

These raw SNP sites were subsequently filtered using GATK VariantFiltrationWalker (DePristo, Banks et al. 2010; McKenna, Hanna et al. 2010). During the calling stage, each variant is annotated with information derived from the BAM files (i.e., depth of coverage, allele balance, the result of a statistical test for strand bias, etc.). Filters were then applied to each of these annotations to reject likely false-positive SNPs. SNPs that met any of the following filters were eliminated from the final call set:

- Allele balance for hets (ref/(ref+alt)) (AB) >= 0.75.
- Length of adjacent, homopolymer run where the base matches the alternate allele of the SNP (HRun) > 3.
- Ratio of discovery confidence to read depth (QD) < 5.0.

These filters were based on comparisons of annotation behaviour for known SNPs (i.e., present in dbSNP build 129) to behaviour for novel SNPs (a mixture of true- and false-positives). Novel SNPs with annotation profiles markedly different from known SNPs tend to indicate false-positive status. A threshold was placed such that it would affect few known sites, while eliminating the offending a subset of the novel SNPs, presumed to be false-positives.

Conditional on a site passing the filters, genotypes on all samples with data were called independently using a flat prior.

5.3.3. Boston College SNP Calls

SNPs were called using the GigaBayes package (http://bioinformatics.bc.edu/marthlab/GigaBayes), which is an extension of the original Bayesian SNP caller PolyBayes (Marth, Korf et al. 1999). The method uses many of the same features as the other methods calling SNPs directly from primary sequence data, described above. First, genotype likelihoods are computed in a way similar to samtools, using a simple model to account for the non-independence of errors, and after excluding reads with mapping quality<20 and sites with base quality <10. Second, the algorithm computes the posterior probabilities for the most likely constellations of sample genotypes by combining the genotype likelihoods with a prior on variant frequency of 0.001/k per site, where k is the number of non-reference alleles in the samples, and assuming Hardy-Weinberg equilibrium. Third, it computes the overall a posteriori probability that the site is variant, by summing over the probabilities of sample genotype constellations containing at least one non-reference allele. Fourth, it obtains the maximum a posteriori estimate of the individual sample genotypes. Candidate SNP sites were filtered so that variant calls had a PHRED Q score of at least 40 and at least one individual with a non-reference variant with genotype quality of at least 10.

5.3.4. Generation of Exon Project SNP Release Set

The SNP release consisted of 7 population-specific call sets. In a given population, the set of SNP sites for the release call set was formed as the intersection of the BC and the BI calls. As for the low-coverage project, the quality of the consensus calls (as measured by transition-transversion ratio, dbSNP concordance and genotype accuracy against external data; data not shown) was better than either individual data set. Genotypes were reported for every sample for which the BC and BI genotype calls agreed. Samples for which the estimated genotypes differed are reported as missing data. The merged call set was not filtered further.

6. Validation of SNP Calls

Targeted genotyping at predicted SNP locations was used iteratively through the pilot phase to estimate both the false positive rate for sequence based SNP discovery and the error rate of sequence based genotyping. This validation was not intended to characterize the final SNP calls, but rather to inform technical and analytical artifacts so that they could be addresse. Moreover, such validation of sites called as polymorphic cannot provide information about the false negative rate for SNP discovery.

The sequence based SNP discovery process deliberately disregards existing information from dbSNP (as of build 129). However, SNPs previously reported as polymorphic in dbSNP naturally have much higher probability of being variable than sites not previously marked in dbSNP, and empirically, show much lower error rates than (novel) sites that were not previously in dbSNP. For this reason, the experimental validation deliberately oversamples from the novel sites.

6.1. Validation of Randomly Selected Low-Coverage SNPs

250 novel (not found in dbSNP build 129) sites were chosen at random from the union of the three preliminary (March 2009) call sets on chromosome 20 in each of the three populations. Sequenom probes were considered to have genotyped effectively if they had less than a 5% no-call rate and a Hardy-Weinberg chi-squared value of less than 3.78.

Validation analysis was conducted in December 2009, using the 2-of-3 intersection SNP call set from March 2009. The true positive (TP) rate was found to be approximately 77%. However, investigation of the SNPs that did not validate revealed a suitable diagnostic for removing a high proportion of the calls that did not validate. Namely, positions with a high proportion of reads with zero mapping quality were highly enriched for SNPs that did not validate. For this reason, the final release set of SNPs has been filtered, as described in section 5.1.4 of the supplementary material.

From the novel SNPs in the final release set, 505 had passed Sequenom design and genotyped effectively. 154 were in CEU, 210 in YRI, and 141 in CHB+JPT. These show true-positive rates for novel sites of 90.3% in CEU, 91.4% in YRI and 84.4% in CHB+JPT. Allowing for the fraction of all sites that were already in dbSNP, this suggests false discovery rates (FDR) of approximately 3.26% in CEU, 4.0% in YRI and 4.3% in CHB+JPT (See Supplementary Table 3).

6.2. Validation of Low-Coverage SNPs with Large Frequency Differences between Populations

SNPs with large frequency differences between populations were also selected for validation. These SNPs were selected randomly from the set of SNPs with a frequency difference of at least 0.6 between two populations, as determined from the merged SNP call set. SNPs were chosen for validation regardless of dbSNP status. Equal numbers were picked from each pairwise population comparison (CEU vs CHB+JPT, CEU vs YRI and CHB+JPT vs YRI).

Of the SNPs that passed the mapping quality filters, 50 SNPs in CEU, 46 in YRI and 41 in CHB+JPT passed Sequenom design and genotyped effectively. The true positive rates were 96.0%, 95.7% and 95.1% for CEU, YRI and CHB+JPT respectively (see Supplementary Table 3).

6.3. Validation of Loss of Function and Non-Synonymous Low-Coverage SNPs

The validation set for loss-of-function (LOF) variants in the low-coverage project included all 50 predicted stop-introducing and splice-disrupting SNPs in the 2-out-of-3 call set on chromosome 20, as well as 32 such variants present in only one call set. Finally, 40 predicted frame-shift-inducing indels called by one of the three indel call sets with the highest quality metrics were also targeted. 130 predicted non-synonymous SNPs were also randomly selected from the 2-out-of-3 call set on chromosome 20. There were dbSNP sites in the LOF and nonsynonymous validation sets, although variants previously seen to be variable in either HapMap or the Illumina 1KG chip were excluded.

Having applied mapping filters, a total of 87 SNPs in CEU, 100 in YRI and 64 in CHB+JPT passed Sequenom design and genotyped effectively. For this set of SNPs, the TP rate was 92.0% in CEU and YRI, and 90.6% in CHB+JPT.

6.4. Validation of Trio SNPs

For the trio project, two independent sets of preliminary genome-wide SNP and genotype calls were made in September 2008 at the Sanger Institute and the University of Michigan, using sequence data from all three members of the CEU trio. 88% of these SNP locations were already contained in dbSNP. A total of 1300 sites were chosen for validation from the union of these two call sets. From the UMich calls, 100 sites were chosen at random from calls already in dbSNP build 129, and 600 sites were chosen at random from calls not in dbSNP. From the Sanger calls, 600 sites were chosen at random from calls putatively not in dbSNP. However, the Sanger calls were apparently matched against an earlier version of dbSNP, since 44% of these 600 sites are in fact found in dbSNP build 129. 1153 sites passed Sequenom primer design criteria and were genotyped on 179 individuals from the low-coverage project plus all 6 individuals in the trio project. After omitting three individuals from the low-coverage project with low Sequenom call rates, 1117 sites have better than 90% call rate. 1064 of these sites are included in the intersection of the two preliminary call sets.

At 308 sites already in dbSNP, 2 are apparent false positives (since the Sequenom genotyping shows all three trio individuals to be homozygous for the reference allele) and 13 more sites show one or more discrepancies between the Sequenom and sequence-based genotype calls, for a total of 26 genotype discrepancies. This estimates a 0.65% false positive rate for SNP discovery at dbSNP sites and a 2.8% rate of genotyping errors. 756 sites not in dbSNP show a 6.3% false positive rate and a 6.8% per genotype error rate for novel sites. However, novel sites are only 12% of the total, so overall, this estimates a 1.3% false positive rate and 3.3% genotyping error rate for the intersection of the two original call sets. This intersection was released in December 2008 as CEU trio SNP calls. The Sequenom data also include 53 sites that were in one of the two call sets and not the other. These show 63% false positive rate and 52% genotyping error rate. Thus, the intersection of the two call sets shows higher accuracy than either call set individually.

We also evaluated the March 2010 final release set against the 2008 Sequenom data, although the results will be slightly optimistic, since information from the 2008 validation experiment may have been used to guide the choice of subsequent filtering criteria. The March 2010 set is more conservative than the earlier release. 286 sites in dbSNP build 129 and 682 sites not in dbSNP show 0.5% false positive rate and 1.5% genotyping error rate overall. Separate estimates for the dbSNP and novel sites in the March 2010 calls show 0.35% and 1.6% false positive rates, and 1.4% and 1.9% genotyping error rates, respectively.

The Sequenom genotyping results also contain genotypes for all three members of the YRI trio. However, the loci being genotyped were ascertained exclusively from the CEU trio, and only 262 loci are called as polymorphic within the YRI trio, with more than half coming from dbSNP. Because of an unknown selection bias for this subset of sites, estimates for the false positive and genotyping error rates would not be reliable.

6.5. Validation of Exon Project SNPs

Three series of validation experiments using Sequenom genotyping were carried out on sites from both primary exon SNP call sets in addition to calls in the intersection. Sites were filtered out of final genotyping results if they exhibit a Hardy-Weinberg equilibrium violation, have high no-call rates across samples or list every sample as homozygous variant.

Random Sampling

126 sites not found in dbSNP 129 were selected at random from the intersection between the BC and Broad call sets. These sites were genotyped using Sequenom in the various populations and the average TP rate was 94%, suggesting an overall average FDR of 1.97%

Population Specific Discovery

Approximately 135 sites not found in dbSNP 129 were chosen in each of CEU, YRI, and CHB+JPT populations, and regardless of the allele frequency were validated via Sequenom. Among these, the observed TP rate was 95% or greater, suggesting an overall FDR of 2.4% or less.

Low-frequency sites across all samples
68 sites at low frequency (35 singletons and 33 at 2 – 5 occurrences) not found in dbSNP 129 were selected from the intersection of BC and Broad calls. These sites were genotyped using Sequenom. All but one of these sites were found to be polymorphic, suggesting an overall TP rate of 98.5%.

The Sequenom validation outcomes for these sets are given in Supplementary Table 3. In all three validation series, the validation rate is over 90%. The raw totals (not weighted by the number of calls in each category) yield an overall validation rate of 91.2%.

6.6. Custom Validation Genotyping Chip

In early 2009 Illumina offered to make a custom genotyping array to validate preliminary variant calls from the project, and in May 2009 a set of 150,000 SNP calls was submitted for chip design as described in detail at ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/supporting/Illumina_Human1KGP-12/chip_annotation/Illumina.1KGP-12.design.summary.pdf. Included were ~55,000 non-synonymous and splice site SNPs in Ensembl genes, ~70,000 SNPs around GWAS signals in the NHGRI table of GWAS hits, and ~10,000 SNPs in ENCODE regions resequenced for the HapMap 3 project.

Approximately 120,000 of these sites converted into working assays, and the resulting chips provided by Illumina were used to genotype all 1000 Genomes Project samples available at the time. The resulting genotypes are available on the project FTP site.

Based in large part on data from the 1000 Genomes Project, including this experiment, other next generation genotyping chips have been designed, including the Metabochip (http://www.sph.umich.edu/csg/kang/MetaboChip/) and the 2.5 M Omni chip (http://www.illumina.com/applications/gwas.ilmn) have been designed.

6.7. Y Chromosome and Mitochondrial SNP Calls

6.7.1. Y Chromosome

Genotype likelihoods on the Y chromosome in the low-coverage data were generated using GLFTools v3 (http://sourceforge.net/projects/samtools/files/glftools, which implements likelihood and variant calling algorithms of Li and Durbin (2008)). Combining these with an expected population-scaled mutation rate θ = 0.001, a heterozygosity prior penalty of 50, and an RMS mapping quality threshold of 60, generated a list of 49,290 candidate SNPs. The phred-scaled heterozygous prior penalty corresponds to a 10^-5 heterozygosity rate; this penalty is designed to eliminate heterozygous calls except at sites with very high evidence for heterozygosity, which can be excluded as mapping artefacts. For sites that were called as variable, genotype posterior probabilities were re-calculated using a uniform genotype prior (theta=1). Following this two sets of filters were applied.

First Stage Filters

- Heterozygosity Filter. Removed sites with heterozygous calls of any quality (taking into account the het-prior), or more than 3 different high-certainty alleles.
- Depth Filter. Removed sites with under 1/2 or over than 3/2 times the mean depth at HapMap sites.
- Proximity Filter. Removed sites that are within 4 bp of other candidate sites.

A total of 5,839 sites passed the First Stage filters. Of the novel sites within the unique Y region that passed First Stage filters, 200 were randomly selected for capillary resequencing. However, 4 failed primer design and 40 primer pairs were excluded as being non-specific by nucleotide BLAST (blastn). The remaining 156 were amplified in two males and a female sample and 117 that gave male-specific amplimers were selected for sequencing. Of these a total of 112 sites produced good sequence traces on either the forward or reverse strand, 69 of which (62%) validated as non-reference.

Following this, Second Stage filters were applied:

- Haplotype Filter. Removed sites that are polymorphic in more than 1 major haplogroup.
- Quality Filter (Singletons). Removed sites with a non-reference quality < 50.
- Quality Filter (Non-Singletons). Removed sites with a mean non-reference quality < 30 and a total non-reference quality < 100.

2,870 sites passed these Second Stage filters, of which 24.1% were in dbSNP 129. Of sequenced Second Stage filtered SNPs, 55/56 validated, giving an estimated False Positive rate of 1.8% (with a 95% upper confidence bound of 8.2%). A final filter was applied that flagged sites that lie outside of the approximately 12 Mb Unique Y-specific Region (UYR), though calls that failed this filter are included in the final call set. A total of 1,971 SNPs fall inside this region, of which 26.0% were in dbSNP.

6.7.2. Mitochondrial DNA Analysis

Mitochondrial DNA (mtDNA) sequence data were examined from 208 datasets from the low-coverage project, each corresponding to a combination of individual and platform. 22 sequences were excluded because of sequencing quality and 23 individuals were represented by two different platforms, so the final dataset was from 163 individuals. Mean coverage for each individual ranged from 37.7X to 3535X.

Reads mapping to mtDNA were extracted from the original bam files and regenerated as mtDNA bam files. For subsequent analysis, we used the SAMtools package (Li, Handsaker et al. 2009) to generate an initial pileup file from which a consensus sequence and variation information for each individual were obtained. Consensus sequences were aligned and compared with the revised Cambridge Reference Sequence (rCRS; Andrews, Kubacka et al. 1999) and initial haplogroup assignments made using a web-based haplogrouping program (Brandon, Ruiz-Pesini et al. 2009).

Manual examination of these initial sequences revealed a lot of mis-aligned reads, and that these reads caused many ambiguous heteroplasmy calls. A java script was used to filter reads based on the NM (number of mismatch) information in the SAM files, removing reads with >10% mismatch (typically 1-5% of initial reads). This filtering step was effective and ambiguous heteroplasmy positions were removed (data not shown).

Mitochondrial DNA has the advantage that the sequence variation falls into a well-defined and (for many parts of the phylogeny) well-understood phylogenetic pattern (Bandelt, Lahermo et al. 2001; Bandelt, Salas et al. 2004). Using this approach, we curated the 163 sequences manually. All positions (54 position/individual combinations) that were ambiguous or surprising according to this analysis were re-sequenced with conventional capillary sequencing and 52 provided an interpretable result that was incorporated into the final dataset.

The 163 sequences studied here were successfully classified into specific (sub-)haplogroups (Supplementary Figure 5). Each continental sample was perfectly divided into population-specific haplogroups as expected (e.g., Haplogroup L for African samples, Haplogroup H for Europeans and Haplogroup D or B for the East Asian samples) according to their previously-identified continental affiliations (van Oven and Kayser 2009).

Heteroplasmies were also assessed in this study (Supplementary Table 10). Heteroplasmy was called with a mean MAF of 26%, with seven calls having a MAF <10%. Heteroplasmies within individuals were first detected in consensus sequences generated from the BAM files after filtering out ambiguous reads. The MAF at each locus was estimated by counting the number of occurrences of each allele in the reads spanning the locus, and confirmed by manual inspection using the program IGV (Integrative Genome Viewer: http://www.broadinstitute.org/igv/).

Of the 163 samples, 85.9% showed at least one heteroplasmic position. Length heteroplasmy was observed in 79% of individuals in the known HVS1, HVS2 and HVS3 C-stretch regions (Supplementary Figure 6A). Point heteroplasmies were identified in 45% of individuals, distributed through the whole mtDNA genome (Supplementary Figure 6B). Irwin et al. (Irwin, Saunier et al. 2009) examined the control region of 5,015 individuals and reported length heteroplasmy in 52%, a little lower than the 79% identified here. Since all the C-stretch regions are located in the control region of the mtDNA, this difference could reflect a small difference in either sensitivity of detection or source material (blood or buccal compared with lymphoblastoid cell line). The same authors reported point heteroplasmy in the control region in approximately 6% of their samples. The incidence of point heteroplasmy in the control region in this study was approximately 4.9%, a similar and indeed slightly lower number, suggesting that somatic mutation during cell culture has not increased the level of heteroplasmy substantially. The higher overall level of point heteroplasmy detected in this study thus most likely reflects the accessing of the complete mtDNA molecule, and the more uniform distribution of point heteroplasmy.

7. Short Insertion/Deletion (Indel) Calls

The process of generating indel variant and genotype calls for the genome-wide data was broken up into three stages: candidate indel identification, calculation of genotype likelihood through local re-alignment, and LD-based genotype inference and calling. This workflow was chosen because, in contrast to SNPs, it is not possible to enumerate all potential indel mutations, and compute the posterior support of each. Instead, a pseudo-

Bayesian approach was used, in which potential candidates are first obtained, and then subsequently tested (together with the reference) in a Bayesian framework.

All likelihood-based calculations were made from the Illumina data, although candidate indels from the 454 and SOLiD reads were considered.

The low-coverage project release consists of indel calls on the finished autosomal sequence. No indel calls on the X, Y, mitochondrial, or unplaced chromosomes have been made.

7.1. Generation of Low-Coverage Candidate Indels

Candidate indels for low-coverage project were obtained using a diverse set of algorithms, working from different primary data sets. The methods were tuned for high sensitivity. Sequenom validation of a small number of randomly chosen calls (50 per set) showed that all candidate call sets were enriched for false positive calls. The 454-based call sets were excluded from the validation experiment because of either low sensitivity or high false-positive rates as indicated by low dbSNP129 concordance. Candidate indels were used subsequently to identify high-quality variant calls (see next section).

The “Broad” method uses the GATK suite of tools to partial re-align reads to identify possible indels directly from read data, using both Illumina and SOLiD reads (DePristo, Banks et al. 2010; McKenna, Hanna et al. 2010).

The “Pindel” method implements a split-read approach; when either side of a single read map confidently to nearby locations on the reference, a potentially large indel is inferred (Ye, Schulz et al. 2009).

The “Oxford” method uses the Stampy read mapper to identify indels by directly aligning reads to the reference. Indels seen in 2 or more samples were reported (Lunter and Goodson 2010).

The “SAMtools” method uses the indel caller implemented in the SAMtools package to call indels directly from 454- or Illumina BAM files (Li, Handsaker et al. 2009).

The “Yale” method uses a split-read approach to identify small to potentially large indels.

7.2. Indel Genotype Calling

The Dindel algorithm (Albers, Lunter et al. 2010) was used to generate both indel calls and individual-level genotype likelihoods. The basic idea of Dindel is to realign all reads mapped to a genomic region to a number of candidate haplotypes. Each candidate haplotype is a sequence of 120 bp that represents an alternative to the reference sequence, and corresponds to the hypothesis of an indel event and potentially other candidate sequence variants such as SNPs. By assigning prior probabilities to the candidate haplotypes, the posterior probability of a haplotype and consequently an indel being present in the sample(s) can be estimated. This Bayesian approach allows us to model different types and rates of error consistently in a single framework. The advantage of modeling hypotheses as candidate haplotypes is that all differences between the read and the candidate haplotype must be due to sequencing errors. Thus, in the realignment of a read to a candidate haplotype, Dindel accounts for the increased sequencing error indel rates in homonucleotide runs, as well as the base-qualities, and naturally separates contributions of errors from statements about biological differences. The process of realigning reads to candidate haplotypes also resolves the issue of mismatches around indel event and corrects alignment artifacts introduced by the read mapper. Furthermore, we deal with mapping errors of by interpreting mapping quality as the prior probability that a read should align to any of the candidate haplotypes (Li, Ruan et al. 2008) which effectively reduces the weight of reads that cannot be confidently mapped to that location in the genome.

Candidate haplotypes are generated by combinatorially combining the candidate indel with three additional candidate sequence variants. The additional candidate sequence variants are generally SNPs but can also be indel variants. Thus, with 4 sequence variants, 16 candidate haplotypes are generated, except when candidate deletions may result in identical haplotypes and reduce the number of unique candidate haplotypes.

First, initial indel calls were made for each indel in the candidate set by jointly analyzing the reads of all individuals in the same population using the Bayesian EM algorithm in Dindel (Albers, Lunter et al. 2010). The Bayesian EM algorithm provides both an estimate of the posterior probability that the candidate indel segregates in the population and an estimate of the population allele frequency, under the assumption that reads were sampled uniformly from the individuals in the population. Here, we also made the assumption that at most one non-reference variant was allowed to segregate per site. While it is expected that this assumption does not hold for all indel polymorphisms, this assumption appeared to lower false discovery rate, and was required for the subsequent genotype imputation stage.

Next, genotype likelihoods were generated for each indel called by the Dindel Bayesian EM algorithm. The genotype likelihoods were computed after the Bayesian EM stage of Dindel, as the genotype likelihoods depend on the estimated frequencies of the variants surrounding the candidate indel for which the genotype likelihoods are desired. The genotype likelihood for the candidate indel takes into account the population frequencies of the additional sequence variants in the candidate haplotypes without assuming linkage equilibrium.

Finally, for low-coverage indels QCALL (Le and Durbin 2010) was used to impute genotypes using the genotype likelihoods for the indels called by the Dindel Bayesian EM algorithm. For a small number of indels called by Dindel (~0.20%), QCALL did not confirm that the indel site was variable (all genotypes were imputed as homozygous reference). These indel cases were filtered from the final call set. For trio indels a Mendelian prior was enforced, essentially identical to that used by Michigan for their trio SNP call set.

7.3. Homonucleotide Sequencing Error Indels

In the approach taken by Dindel, all differences between the read and the haplotype are assumed to be the result of sequencing errors. Dindel accounts for increased sequencing error indel rates in homonucleotide runs. The sequencing error rate is a parameter in the read-to-haplotype probabilistic realignment model that depends on the length of the homonucleotide run in the candidate haplotype.

We estimated the probability of observing a sequencing error indel in a homonucleotide run as a function of the length of the run from the low-coverage project data. The details of this error model are given in the Dindel paper (Albers, Lunter et al. 2010).

7.4. Indel Filtering

Dindel considered both mapped reads and unmapped reads of which the mate is mapped within a distance of the mean library insert size plus 4 standard deviations of the candidate haplotype. Only reads with mapping quality >0 were realigned and used to make calls. Since Dindel uses mapping qualities in the probabilistic model to weight reads, and the reads with high mapping quality dominate the inference.

Dindel and QCALL estimate for every candidate indel the posterior probability that the variant segregates in the population. In addition to requiring that the posterior probability estimated by both Dindel and QCALL should exceed 90% (q10), the following filters were applied post-calling. We required that the estimated population allele frequency >= 1/number of reads, and that at least one read mapped to the forward strand and one mapped to the reverse strand cover the indel. Furthermore, if more than one candidate indel had posterior probability >90% in a window of 30 base pairs, only the one with highest posterior probability was called. We only called indels in windows where the number of reads aligned to the haplotype window was in the 2nd-99th percentile range. Finally, indels in homonucleotide runs longer than 10 nucleotides were filtered.

7.5. Validation of Low-Coverage Indel Calls

Novel (not found in dbSNP build 129) indels were chosen from chromosome 20 and genotyped via Sequenom. Some of the indels chosen appeared in multiple call sets. The criteria for inclusion in validation were:

- Homopolymer or tandem-repeat context of at most 10 bp long
- Indel length at most 50 bp

79 indels in CEU, 59 in CHB+JPT, and 152 in YRI designed and genotyped effectively using Sequenom. Of these, the observed TP rate was 98.7% for CEU, 99.3% for YRI and 94.9% for CHB+JPT (Supplementary Table 3).

7.6. Generation of Trio Project Indel Calls

The process of generating indel calls for the two trios was broken up into analogous stages to those in low-coverage indel calling: candidate indel identification, calculation of genotype likelihoods through realignment, and trio-aware indel calling. All the components used for these steps have been described previously; only their combination is novel.

7.6.1. Generation of Candidate Indels

The trio project indel calls were made from the Illumina data only. As candidates we took all gaps identified by the read-mapper MAQ across all six members of the two trios. Thus, candidates from the YRI trio were tested in the CEU trio and vice versa. In total 14.8M candidates were identified.

7.6.2. Indel Genotype Likelihoods

Genotype likelihoods were calculated for each individual for each candidate indel by Bayesian realignment of reads to candidate haplotypes using Dindel (Albers, Lunter et al. 2010). The main difference with the genotype calculation for the low-coverage project is that here the prior probability of each haplotype pair based on the variants contained in the haplotypes was used, rather than the estimated population haplotype frequencies.

7.6.3. Trio-Aware Indel Calling

Indel calls were made in each trio by assuming Mendelian segregation of the candidate indel. As a result, *de novo* mutations are not included in the call set. The posterior probability for the genotype configurations of the trio members was calculated assuming a Mendelian allele transmission model without mutation, and a simple prior probability distribution over the genotype configurations of the parents, where the distribution over genotypes with at least on non-reference allele was uniform with mass 1/10000. The model allowed for multi-allelic indels. An indel was called if the posterior probability that at least one individual had a non-reference allele was >99%.

7.6.4. Indel Filtering

We applied most of the filters that were also applied for the low-coverage project indel call set. We required that the indel was covered by one read mapped to the forward strand and one read mapped to the reverse strand, and reported only the indel with highest posterior probability in the window of 30 bp. We only called indels in windows where the number of reads aligned to the haplotype window was in the 2nd-99th percentile range. Indels in homonucleotide runs longer than 10 nucleotides were filtered.

7.7. Generation of Exon Project Indel Sites and Genotype Calls

Indel calls for the exon project were derived from the union of three primary call sets. The indel calls were produced using the Illumina platform at Baylor and Broad and the Roche 454 platform at Baylor, in three independent pipelines. The union includes all 697 Exon Pilot individuals from 7 populations (90 CEU, 109 CHB, 107 CHD, 105 JPT, 108 LWK, 66 TSI and 112 YRI).

The Broad indel calling pipeline ran the GATK Indel realigner at the population level, arranging alignments in a consistent fashion across samples. Following this cleaning, a heuristic method (at least ~40% of reads in a sample have a consistent, consensus indel) was applied to the realigned reads. Only variant sites were identified; specific genotypes were not called. These variant sites were identified per sample, and then aggregated to make population-level calls.

Variant sites were filtered via a simple additive points system (sites with fewer than 2 points were filtered out):

Consensus mismatch rate < 1.0 (3.0): +2 (+1) points
Consensus mismatch rate < reference mismatch rate: +1 point
Homopolymer run <= 2: +1 point

Strand bias >= 0: -1 point
Homopolymer Run >= 5: -1 point
Single base event: -1 point

The Baylor indel calling pipeline used a logistic regression model of pertinent variables implemented in Atlas-INDEL2 (Challis et al. in prep) to make indel calls for individual samples from the MOSAIK alignments described above. This was followed with a genotyping step, so for each sample-site if the 'variant call reads/total coverage' was >= 0.2 and < 0.9, the sample was considered a heterozygote. Otherwise it was classified as a homozygote.

The Roche 454 data was processed using a separate pipeline by the BCM-HGSC. Indels were called using Atlas-Indel (Wheeler, Srinivasan et al. 2008) and subjected to the following filters - reads with high substitution or indel rates are filtered; reads where the indel was near the 3' end are filtered; indels that could have resulted from flow errors are filtered; at least 5% of the indel's read strands must be in each direction; singletons are filtered; and 2 base-pair indels with an allele count of less than 5 are filtered.

The union of the three independent call sets that fell within the consensus target regions formed the release call set. In merging call sets and calculating concordance, indels were considered equivalent if they were within 5 base-pairs of each other and of the same type. Equivalent indels were merged, so that no indel was counted more than once.

8. Structural Variation

Several methods were used to discover structural variants (SVs) from trio and low-coverage raw sequencing reads. Algorithms underlying several distinct rationales for SV detection have been used in the SV discovery process. The high-confidence SV set reported in this study is based on validated SVs and on such discovered by algorithms achieving an FDR ≤10% (see Supplementary Tables 4A, 4B). Further details on SV callset generation and analyses on the full set of SVs generated by the 1000 genomes project structural variation analysis group are described in a companion paper.

8.1. Discovery of Structural Variants in the 1000 Genomes Pilot Project

8.1.1. Paired-End Mapping

The paired-end mapping approach utilizes paired-end sequencing reads and identifies SVs by relating the genomic mapping positions of both ends of a pair-end (also called *read-pair*, and, depending on the library preparation protocol, *mate-pair*) to the paired-end insert size distribution. A number of different approaches were used for mining paired-end insert sizes and inferring SVs based on the identification of deviations from the expected insert size, as described in the following. These approaches usually employ additional criteria for discerning high-confidence SV assignments from such made at low confidence.

8.1.2. Deletion and Insertion Analysis with the AB Large Indel Tool

The AB large indel tool infers insertions and deletions by identifying positions in the genome in which the pairing distance between mapped read-pairs (mate pairs) is significantly deviated from what is expected at the given level of span coverage (i.e., physical coverage in terms of read-pairs spanning the genome). A look-up table is created in which the amount that the read-pairs must be deviated to achieve one standard deviation of significance is the standard error at each level of physical coverage. This produces an asymptotic curve in which the minimum size of detectable indels at a given level of significance drops rapidly as the span coverage increases. The look-up table is used to determine the significance of the deviation in average insert size at each position in the genome. Regions of the genome that are significantly deviated are selected as candidate indels. Hierarchical clustering is used to segregate candidate indels into groups. Clusters with <2 read-pairs are removed, and each cluster is assessed to differentiate homozygous and heterozygous SVs (McKernan, Peckham et al. 2009), i.e., if 2 different clusters of pairing distances are observed, the SV is called heterozygous; if there is only one cluster, it is called homozygous. The tool can be downloaded at SOLiD software tool website.

8.1.3. Deletion and Insertion Analysis with PEMer

The PEMer tool, which identifies SVs by mining paired-end mapping data (Korbel, Urban et al. 2007), was applied to 454 paired-end and SOLiD mate-pair reads. The PEMer calling method involves a number of subsequent steps, including read-mapping onto the reference genome (using megablast for 454 reads and Corona Lite for SOLiD reads), removing duplicates, SV-calling with PEMer, and quality checks to generate high-confidence calls. The PEMer pipeline for calling SVs contains the following steps, which were carried out as described previously (Korbel, Abyzov et al. 2009): identification of outliers whose paired-end alignment distances on the genome deviate significantly from expected distances (with cutoffs determined based on the paired-end cluster size), outlier clustering, and cluster merging. For SV calling with 454 data, a p-value cutoff (Korbel, Abyzov et al. 2009) of 0.05 was applied. For SV calling with SOLiD data, several post-processing steps were applied to compensate for the placement of short reads, which is more challenging that long-read placement: 1) consistency check of span-size of each read-pair within cluster (only read-pairs whose span-size is within 15% deviation from the median of span-size in each cluster are kept); 2) recovery of sub-clusters from a big cluster that better define the boundaries of an SV event; 3) and span size check of the innermost reads within a cluster against the SV event size (the distance between the two innermost reads supports the SV event). Sub-cluster recovery was made to clusters violating the condition that the rightmost read on the left is left to the leftmost read on the right, by identifying sub-clusters that meet the condition. SV size was estimated as a difference between the average distance of all read pairs in a cluster, or sub-cluster, and the average insert size.

8.1.4. Deletion Analysis with BreakDancer

We ran BreakDancer (Chen, Wallis et al. 2009) to identify deletions from Illumina paired end data in both the trios and the low-coverage project. In each library, paired-ends with MAQ (Li, Ruan et al. 2008) mapping qualities greater than 35 were tagged as discordant if their outer distances were larger than the mean plus four standard deviations of the insert size. We then searched for genomic regions that anchor significantly more discordant paired-ends than expected on average. Data from family trios was jointly analyzed together with low-coverage data from the same population. A putative deletion is derived from the identification of one or more regions that are interconnected by at least two discordant paired-ends. At each deletion locus, individual genotypes were determined by analyzing the supporting discordant paired-ends. The start and the end coordinates are defined as the inner boundaries of the constituent regions that are closest to the suspected breakpoints, while the size is estimated by subtracting the mean insert size from the average spanning distance in each library and then averaged across libraries.

8.1.5. Deletion Analysis with VariationHunter

All high-quality reads (average phred score ≥ 20) were mapped to reference human genome build 36 with edit distance ≤ 2 with the mrFAST algorithm (Alkan, Kidd et al. 2009). Paired-end sequences mapped with a span within the size threshold (average± 4×stdev) were returned as concordant. ; All locations for non-concordant (i.e., discordant) paired-ends returned by mrFAST as potential mapping location were considered for SV detection. The SV detection algorithm VariationHunter (Hormozdiari, Alkan et al. 2009), uses a maximum parsimony optimization function to minimize the total number of structural variants in a given genome while assigning paired-ends to a single location. To achieve this goal, VariationHunter first creates all maximal valid clusters (Hormozdiari, Alkan et al. 2009) of discordant paired-ends. Maximal valid clusters are defined as maximum number of paired-ends that support the same potential variation, where no other paired-end can be added to such clusters without conflicting the class and size of the structural variant signaled by the other paired-ends in the cluster. VariationHunter then uses an approximate solution to the set-cover problem to find the minimum number of total structural variations.

8.1.6. Deletion Analysis by Paired-End Mapping at WTSI

Reads from the Illumina/Solexa platform were mapped with MAQ (Li, Ruan et al. 2008), then insert size distributions were analyzed for each library separately on chromosome 20. To select aberrant paired-ends from each library a cut-off was determined by using a drop in the density function in the relevant insert size distribution. Then paired-ends with mapping quality of at least 20 were ordered by start position and two aberrant pairs were allocated to the same cluster if their start positions were not further apart than a given threshold (ten times the median absolute deviation of the insert size distribution) and similar their end positions. In a second step neighboring overlapping paired-end clusters were merged. To obtain a final set of putative deletions, the candidate clusters were filtered by the product of the number of paired-end reads per cluster and their mapping quality. Additionally they were filtered by the length of the cluster interval of the first or forward reads and the same filter was used for the second or reverse reads (less than ten times the maximum median absolute deviation of all libraries). Finally, deletion candidates were also filtered by deletion size. A deletion was called if the innermost reads of its defining cluster were further apart than the median of all the cut-offs used for extracting aberrant paired-ends in the different libraries. Deletions greater than 1 Mb were discarded.

8.1.7. Mobile Element Insertion Analysis with SPANNER

Transposable elements were identified by two approaches, a paired-end based approach applied to Illumina paired-end data and a split-read based approach applied to 454 data (described below). Further details with regard to mobile element detection and analysis are provided in a companion paper.

Alignment. All SPANNER calls were based on MOSAIK (Marth 2010) alignments. MOSAIK version 0.9.1176 was used to align the Illumina and 454 reads from the 1000

Genomes Project low-coverage and trio data against the combined reference genome (build hg18) and mobile element consensus sequences, to enable an analysis of deletions, tandem duplications (see below), and as well as mobile element insertions.

For the Illumina paired-end data the MOSAIK alignment parameters were: (1) maximum mismatch threshold of 4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer; (2) hash size of 15; (3) Smith-Waterman bandwidth of 17; (4) alignment candidate threshold of 25 bp; (5) local alignment search radius of 100 bp; (6) hash position threshold to 100.

For 454 data the parameters for initial paired-end alignments were: (1) maximum mismatch percent threshold of 5%; (2) hash size of 15; (3) Smith-Waterman bandwidth of 51; (4) alignment candidate threshold of 55 bp; (5) hash position threshold to 200; (6) homo-polymer gap open penalty of 4.

*Mobile element insertion detection with Illumina paired-end reads***.** Mobile element insertions were detected as clusters of paired-end fragments in which one end aligns uniquely to the genome and the other end maps to a mobile element reference sequence. All "proper-pairs" in which the mapping distance between the pairs is consistent (*P*-value<0.99) with the paired-end insert size distribution were removed from consideration as supporting fragments for candidate insertions, since these fragments are consistent with already annotated mobile elements. Paired-ends were classified by mobile element type (*Alu*, L1, or SVA) and the orientation of the uniquely mapped end. Supporting paired-ends spanning into the mobile element from the 5' direction (F) or such spanning into the mobile element from the 3' side (R) were selected for clustering. With this convention, an insertion can be identified by two matched clusters of paired-ends spanning into the insertion from both 5' and 3' sides. The same clustering algorithm was used for grouping supporting paired-ends bracketing the insertion, except that the two dimensional clustering space consists of the paired-end orientation (F or R) instead of the event length that was used for deletions and tandem duplications. A minimum of 4 supporting paired-ends (two from each side) was required for each insertion candidate.

Following SPANNER processing, candidate mobile element insertions were filtered with post-processing selection criteria. In addition to “alignabilty” and VNTR masking criteria, two additional criteria were applied:

- Insertion candidate coordinates were compared with annotated loci of *Alu* subfamily, L1 family, or SVA elements. Event loci that map within 400 bp of a corresponding type mobile element annotation were discarded.
- The gap between the F and R clusters outside of the range (-30 bp < gap < 500 bp) were discarded.

8.2. Read Depth Analysis

Read depth analysis identifies SVs through relating the relative depth-of-coverage of genomic windows to a model of the expected depth-of-coverage. Groups contributing to the analysis used different mapping tools and analysis algorithms to segment the raw sequencing data into deletions and duplications.

8.2.1. Read-Depth Analysis with CNVnator

The main step for CNV discovery by CNVnator (Abyzov, Urban et al. 2010) is the partitioning of the read-depth signal into regions with different copy-number. Read-depth was estimated based on reads mapped with MAQ. The read-depth partitioning procedure is based on image processing technique, known as mean-shift-theory (Comaniciu and Meer 2002; Wang, Abyzov et al. 2009). In relation to processing of read-depth data the theory can be applied as follows. A diagram displaying the read-depth signal along a chromosome represents an image that needs to be processed with the aim of identifying different copy number levels. The read-depth signal is proportional to underlying genomic copy number level, but fluctuates around an average value owing to noise. Thus, statistically we can formulate this problem as finding a Probability Distribution Function (PDF) from observed read-depth data, where the PDF itself represents an unknown mixture of PDFs/modes, corresponding to different genomic copy number levels. The density maxima in the distribution of intensities are the modes of the PDF, where the gradient of the estimated PDF are zeros. The mean-shift method presents a way to locate these density maxima without having to estimate the density directly[3]. The mean-shift vector always points in the direction of maximum increase in the density. The mean-shift process is an iterative procedure that shifts each data point to these density maxima.

In short, for each bin in genome a mean shift vector is estimated by comparing the bin's read-depth signal with those in the local neighborhood. The vector points in the direction of bins with most similar read-depth signal, thus effectively segmenting the read-depth signal diagram into local modes of attraction. Boundaries of genomic segments are identified by finding consecutive pair of bins with mean-shift vectors of opposite signs. Afterwards, smoothing of the read-depth signal is performed by averaging signal values within each segment. Note that the mean-shift technique does not require prior knowledge of the number of segments or assumptions about probability distributions. This approach performs the discontinuity preserving smoothing on the read-depth signal through kernel density estimation and the mean-shift computation.

8.2.2. Read-Depth Analysis with mrFast

Read depth analysis with mrFAST comprised the following steps:

Repeat masking the genome: Human reference genome build 35 was first masked using the RepeatMasker tool with the sensitivity option (-s) enabled. We then ran Tandem Repeats Finder to mask tandem repeats shorter than 500 bp.

Read Mapping: All reads were mapped to repeat masked reference human genome build 35 with edit distance ≤ 2 with the mrFAST algorithm (Alkan, Kidd et al. 2009) (note that the resulting SVs are reported as in terms of build 36 / hg18; see below).

Read depth: The human reference genome was first repeat-masked to remove common and tandem repeats. Using mrFAST (Alkan, Kidd et al. 2009), whole genome shotgun reads were then mapped to a set of BACs with known copy-number to determine the average read depth and standard deviation in unique intervals (5 kb of unmasked sequence) over the autosomes and chrX. All reads were then mapped to the reference genome (build35) and deletions were identified by where at least 6/7 consecutive 5 kb windows showing reduced read depth (>average-2stdev). We defined candidate reduced depth of coverage deletion intervals as those intervals > 20 kb in size and with < 70% repeat masked. Finally the build 35 coordinates were lifted over to build 36 using the UCSC liftOver tool.

8.2.3. Read-Depth Analysis by Event-Wise-Testing

Preprocessing of mapped reads. For each sample, alignment (BAM) files were parsed out using SAMtools, and reads of low mapping quality (<Q30) were filtered out. Note that we included multiply mapped reads with a mapping quality of 0 in our analysis to detect duplications. Read-depth was measured by counting the number of mapped reads in 100 bp windows, assigning each read once by its start position. Then, we adjusted the 100 bp window read counts based on the observed deviation in coverage for a given G+C percentage. Our subsequent analysis was carried out on such GC-corrected read counts.

Event detection and copy number estimation. We applied event-wise-testing (Yoon, Xuan et al. 2009) to the 100 bp window read counts per chromosome. The event-wise-testing approach, based on significance testing, rapidly searches the genome for deletions and duplications based on assessing whether criteria of statistical significance are met. Since the number of iterations in event-wise-testing is far less than the number of windows (e.g., 19 iterations for all 2.4 million 100 bp windows on chromosome 1), an exhaustive, fast, and robust search of events on very large data sets can be performed.

Additional filtering criteria were applied to a set of calls as follows: First, clusters of small events (within 500 bp) with a copy-number change in the same direction were merged. Then events with a low absolute difference from the average read-depth, that is, a median read-depth of between 0.75 and 1.25 times the overall mean, were filtered out. Then we tested the significance of each merged event by performing a one-sided *z*-test using a significance level at $P<10^{-6}$, which was deemed to be adequate based on manual inspection of many events at all significance levels. Finally, events of size less than 1 kb, and 2 kb, were filtered out in trios and low-coverage samples, respectively.

Lastly, the copy number of each event in the filtered call set was inferred by rounding the average normalized read counts to 2 (1 for X and Y chromosome in male) dividing by the average of the chromosome in each individual to the nearest integer.

Pairwise comparison of read-depth among individuals. We conducted a comparison of deletions and duplications among multiple individuals. Many of the deleted or duplicated regions in our filtered call set clearly differed in copy number among the individuals. Therefore, we sought to distinguish regions that were polymorphic from those that were monomorphic. For each region called based on event-wise-testing in a given sample, we compared the read normalized counts of 100 bp windows in the region between that sample and each of the other samples by *t*-tests. Deletions and duplications were identified based on the *t*-test *P*-value adjusted for multiple testing and the absolute difference between median read counts (D). Events where at least one comparison had significant p-values and D > 0.5 were designated as polymorphic, the remainder, as monomorphic.

Detection of common deletions based on cross-sample analysis. For low-coverage genomes with mapped coverage greater than 1.2x, the normalized read-depth was calculated as above. Then we calculated the Pearson correlation coefficient of each genomic window with the next one and segmented the series of correlation coefficients to detect highly correlated regions. For each region detected, samples with median read-depth (RD) RD < 1.25, and RD > 1.75 were typed as “deletion” and “normal”, respectively, and ones in between were excluded to avoid ambiguity. Regions with more than 4 occurrences, each, of “normal” and “deletion”, were identified as common deletion regions.

8.2.4. Read-Depth Analysis at Albert Einstein College of Medicine

We aimed at discovering common deletions and duplications in low-coverage data by identifying regions in the genome whose read counts across 162 individuals displayed anomalous patterns of mixtures of (over-dispersed) Poisson distributions. For a given window of 500 bps, we constructed a statistic based on the quasi-likelihood of a generalized Poisson distribution:

$$Q = \sum_{i=1}^{n}\left(y_i \ln \frac{\mu_i}{y_i} - \mu_i + y_i \right) \Big/ \phi_i$$

in which y_i is the observed read count of the i-th individual (after mapping with MAQ and adjustment for GC content), and μ_i and ϕ_i are the mean and the over-dispersion parameter of the Poisson distribution, empirically estimated for each individual. The reference distribution of the statistic Q under the null hypotheses of no structural variant (i.e., everyone has copy-number two) is empirically determined by randomly sampled read counts from other regions of the same 162 individuals. We performed the statistical test along the genome on non-overlapping 500 bp windows and retained those with *P*-values less that 10^{-5}, then implemented several filters to further trim down the list. Finally, we merged neighboring statistically significant windows.

8.3. Split-Read Analysis

Split-read analysis involves the alignment of raw DNA reads onto candidate SV breakpoint junctions by following a gapped sequence alignment rationale. The approach therefore identifies SVs at nucleotide resolution. We used different variant split read analysis approaches (as described below) and identified deletions and insertions using such rationale.

8.3.1. Detecting Deletions with Pindel

Previously we developed the so-called Pindel (Ye, Schulz et al. 2009) method that can identify breakpoints of short indels and SVs, i.e., deletions (1 bp - 50 kb size range) and insertions (1-20 bp size range) from short (36 bp) paired-end reads. We examined the MAQ-based BAM files to select those paired-ends for which only one end can be mapped. The mapping quality of the mapped reads must be larger than 0. The Pindel program uses mapped reads to determine their anchor point on the reference genome and the direction of the respective unmapped ends. Knowing the anchor point, the direction to search for the unmapped read and the user defined maximum deletion size (50 kb), a sub-region in the reference genome can be located, where Pindel will break the unmapped reads into 2 (deletion) or 3 (short insertion) fragments and map the two terminal fragments separately.

In this study we focused on identifying deletions with a size range of 50 bp to 50 kb. We developed a procedure based on the Pindel algorithm to process multiple samples. We added tags to each sequence read to indicate its source (i.e., sample of origin). Then we ran Pindel using the entire pool of reads as the input. We modified our Pindel program to report the sample sources of the supporting reads for each identified indel event. This enabled us to identify the sample of origin for each identified indel. We also adapted the algorithm to include detection of deletions with small insertions of non-template sequences at the breakpoint, and to report a confidence score that is monotonically related to the false discovery rate.

8.3.2. Split-Read Analysis of Deletions and Insertions at Yale

After BLAT sequence alignment and placing 454 reads at their most likely locations in the reference genome, split-read analysis was carried out to search these locations for insertions and deletions in the sample genome by identifying reads that encompass SV breakpoints. To find deletions in the sample genome, we searched for reads that when aligned to the reference genome split on the same strand of a chromosome. To find small insertions that are fully included in the reads, we search for reads whose terminal sequences can be aligned next to each other on the reference genome. For large insertions, we looked for their boundaries, which are found in reads that – except for one of their ends – can be aligned to the reference genome continuously in one block. We scored each of the many initial mappings with an assessment strategy designed to take into account both sequencing and alignment errors.

8.3.3. Split-Read Analysis of Transposable Element Insertions with MOSAIK

Unaligned reads from the initial set of MOSAIK alignments of low-coverage and trio DNA reads were re-aligned allowing for partial read mapping to the mobile element consensus sequences using MOSAIK. The following MOSAIK alignment parameters were used: 15 bp hash size, alignments with more than 5% of mismatch bases were filtered out, and at least 40 bp of the read must align to one of the mobile element consensus sequences. For each re-aligned read, a target region was created, which was defined as the largest region spanning multiple overlapping alignments. Reads were discarded if multiple, non-overlapping alignments were found. The remaining sections before and after the target region are compared. The longer section is kept for further processing, while the other section of the read is trimmed away. If less than 40 bp remained after trimming, the read was discarded. Trimmed reads were then aligned to the reference genome (hg18). The longest alignment was selected if it was at least 5 bp longer than the second longest alignment, otherwise the read was discarded. The remaining reads were considered candidate fragments that support a mobile element insertion event. All mobile element insertion candidates were checked for novelty by aligning the entire read back to the reference genome using relaxed alignment parameters (up to 9% of the read bases could be a mismatch and the aligned length could vary from 90 – 100% of the read length). All candidates that were aligned using the relaxed parameters were discarded from the detection pipeline.

Note that there can be up to three unaligned regions in each candidate: a gap occurring at the beginning of the genomic hit region but not immediately adjacent to the mobile element (genomic gap); a gap occurring within the mobile element (mobile element gap); and a gap occurring at the vicinity of the mobile element and the genomic hit region (mid gap). Candidates with a mid gap longer than 6 bp were discarded. Candidates with a genome or mobile element gap longer than 6 bp were also discarded, except when the target region contained the entire mobile element. Candidates were discarded if any of the following criteria were true: the alignment quality score was less than 40; the mobile element alignment length was less than 60 bp; candidate occurred within 100 bp of annotated *Alu*, L1, and SVA mobile elements.

8.4. Sequence Assembly

The SV group furthermore applied several assembly-based approaches to identify SVs, including deletions and insertions at nucleotide resolution.

8.4.1. Identification of Insertions, Deletions and Complex Events with Cortex

Structural variant calls were made using Cortex, a de Bruijn graph variant caller. This was done by using trio reads, focusing on NA12878, to *de novo* assemble the Illumina and 454 data. Errors were removed from the graph, and then two different algorithms were used to call variants. Both algorithms involved fully assembling both alleles of a variant, including the flanking regions. The first algorithm (“bubble calling”) carried out pure *de novo* assembly of heterozygous sites by looking for motifs in the graph – detecting insertion, deletion and complex variants (involving SNPs and indels in close proximity) of up to 1 kb in size. The second algorithm (“reference assisted”) draws the human genome reference in a different ‘colour’ onto the de Bruijn graph assembly and looks for differences between the individual and the reference, and detected homozygous variants ranging up to 40 kb in size.

8.4.2. Targeted Sequence Assembly using the TIGRA Assembler

The TIGRA assembler was applied to deletion calls in low-coverage SV discovery callsets based on paired-end mapping (SPANNER, Genome STRiP, WTSI’s paired-end mapping caller, BreakDancer, VariationHunter, and PEMer) as well as on two split-read analysis approaches (Pindel and Yale split reads).

For each putative SV in each call set, Illumina reads falling into the interval [START_INNER – 500 bp, START_INNER + 50 bp] and [END_INNER - 100 bp, END_INNER + 450 bp] were obtained from the BAM files of the predicted deletion-containing samples using SAMtools. Each set of reads was fed into the WashU TIGRA assembler and its coupled variant calling pipeline. Cross-match was used to align contig sequences generated by TIGRA to intervals in the reference sequence (hg18) corresponding to the predicted variant and flanking sequence regions of 500 bp on either side of the variant (using the inner start and end coordinates of the variant as starting points). Structural variants were then called on the basis of the pair-wise alignment results. A structural variant was assumed as "confirmed" if the assembly based call was consistent with the original call under the following heuristic criteria: (1) same type (e.g., deletion); (2) assembly derived event size >= 50 bp; assembly derived event size differ from original by <100 bp for the Illumina calls and <500 bp for the SOLiD paired-end calls. To assess the approach we obtained a set of 163 known deletions sequenced in the respective individuals, or inferred at nucleotide-resolution using a genotyping approach, and confirmed 151/162 (93.2%) of these deletions.

8.4.3. Detection of Novel Sequence Insertions From Orphan and One-End Anchored Inserts

To detect novel sequence insertions of at least 200 bp in size, we utilized the “orphan” (both ends unmapped) and one-end anchored (OEA) inserts returned at the initial read-mapping stage. Our novel sequence detection pipeline NovelSeq (Hajirasouliha, Hormozdiari et al.), that was applied for this purpose, works as follows. We first *de novo* assembled the orphan read pairs using AbySS (Simpson, Wong et al. 2009) and screened the resulting contigs for contamination. We removed such contigs that are found to include sequence from the Epstein-Barr virus and other contaminants using MegaBLAST and the NCBI nr database. The OEA reads were then clustered to find a minimum set of reads that signal a sequence insertion, and the unmapped ends of OEA clusters were merged with the orphan contigs to anchor the novel sequence insertions.

8.4.4. Detection of Novel Sequence Insertions with SOAPdenovo

Sequence assembly. We performed *de novo* assembly with the SOAPdenovo algorithm (Li, Zhu et al. 2010). Since sequencing depth of each individual genome is not sufficient for achieving a proper assembly, we pooled the individuals into two groups (i.e., the CEU trio and the YRI trio) and assembled consensus sequences for each group. Consensus sequences were merged into one non-redundant dataset. Through comparison against the NCBI reference genome (build 36/hg18), we identified deletions and novel sequence insertions that were not present in the reference genome

Identification of structural variations. We pre-aligned all assembled scaffolds to the reference genome by BLAT v. 30 with –fastMap and –maxIntron=50 option enabled (Kent 2002). Scaffolds that pre-aligned to identical chromosomes were grouped as scaffold sets. These sets were aligned to corresponding chromosomes by a modified version of LASTZ (Harris 2007) based on V1.01.50 with the following options enabled: high-scoring segment pairs (HSP) chaining option, ambiguous 'N' treatment, and gap-free extension tolerance up to 50 kb. Scaffolds with no hit during pre-alignment were aligned to the reference genome with the same options, and inaccurately predicted gaps in assembly and misalignments were corrected. Best hits were further confirmed using the dynamic-programming algorithm based utility "axtBest" (Schwartz, Kent et al. 2003). Finally, we reported hits that contributed most to the collinearity between a scaffold and a chromosome and required two or more alignments to overlap at the same locus in a chromosome. These alignment best hits with gapped extensions included insertions and deletions.

8.5. Integrative Discovery Approaches Using Multiple Features of Sequence Data

The SV group also applied approaches using combinations of some of the rationales described above, i.e., paired-end mapping and read-depth analysis for SV discovery.

8.5.1. Deletion Analysis with the Genome STRucture in Populations

Genome STRiP integrates diverse features of NGS data (i.e., paired-ends, read depth, and the distribution of evidence across samples and within a genomic locus) to identify genomic loci where multiple features of the NGS data coalesce around a model of alternate structural alleles segregating in a population.

Deletion discovery was performed on the Illumina data for low-coverage samples. Duplicate reads were removed, using Picard MarkDuplicates. In the paired-end component of Genome STRiP, data for all low-coverage samples was pooled, and an initial set of candidate deletion sites was identified as clusters of paired-ends (N >= 2) having normal orientation and an apparent insert size greater than the median of the insert size distribution (for that library/lane) plus ten times the median absolute deviation of insert size from the median for that library/lane.

A set of integrative and population-genetic criteria were then applied to the same data. The ideas in this approach are described in detail in a companion manuscript, but at a high level the following features of the data were recognized: coherence of the aberrantly mapping paired-ends in a cluster around a potential alternative molecular structure that could give rise to all of the paired-ends in that cluster; non-uniform distribution across samples of evidence for an alternative allele; and replacement of a reference allele with a deletion allele, as evidenced by local, sample-specific loss of sequence coverage depth.

8.5.2. Deletion and Tandem Duplication Analysis with SPANNER

Integrative deletion discovery with SPANNER was carried out as follows. The fragment length distribution for each run was created in a scan pass prior to detection so that each paired-end could be classified according to it's length and orientation. Duplicate read pairs were removed from the detection sample and identified as multiple fragments with the same mapped coordinates at both ends. Only paired-ends with both mapping quality values (phred convention of -10log(p)) of at least 30 were considered as supporting fragments. Paired-ends for which the mapping distance between the pairs was at the high extreme (P-value<0.04% in the trios and p-value<0.02% in the low-coverage data) of the corresponding library insert size distribution were considered as discordant. Supporting paired-ends were then grouped using a nearest neighbor clustering algorithm (Knuth 1968; Youssef 1987) such that paired-ends within a group support a given deletion breakpoint. Clustering was done in the two-dimensional space of genome position of the leftmost end of a fragment, and the fragment length. The clustering neighborhood scale for each fragment was set based on the empirical library paired-end insert size distribution. In this way supporting paired-ends from a wide range of library fragment lengths could be clustered without the assumption of a common neighborhood scale. Each paired-end could belong to at most one cluster; ambiguities in the clustering were resolved by selecting the cluster with the highest paired-end density among all possible clusters in the neighborhood of a fragment. Clustering was done on the pooled set of all samples within a data set (trio and low-coverage data were processed separately) and only clusters with three or more supporting paired-ends were considered as candidate deletions. The minimum deletion size was 50 bp. The candidate deletion selection criteria included the following additional conditions to reduce artifacts arising from potential read misalignment:

- "Alignability" in the clustered regions > 0.01. Here "alignability" in a region is calculated as the ratio of non-multiply mapped read coordinates to all coordinates in the clustered regions before and after a deletion breakpoint. This condition served a similar purpose to the mapping Q0 criteria imposed for SNP discovery in low-coverage data.
- Net read coverage over all samples < 2.5 times the expected coverage. The expected coverage was calculated by random sampling 100 regions in the chromosome of the same length and repeat content of a given candidate deletion. Candidate deletions in regions of read depth pileup were discarded as artifact.
- Deletions with overlapping boundaries were consolidated into one deletion. The longer length deletion was discarded as alignment artifact.
- Candidate deletions within 100 bp and a reciprocal overlap > 25% of a candidate tandem duplication were discarded as alignment artifact.

- Candidate deletions overlapping with segmental duplication annotations (UCSC browser “genomicSuperDups“), or with annotated VNTRs (UCSC browser “Simple Repeats” located by Tandem Repeats Finder) were discarded.

Of note, the paired-end detection method brackets the position of the deletion somewhere within the gap between the clusters of paired-ends. This results in an estimate for the deletion coordinates with a corresponding uncertainty. The breakpoint resolution is limited by the combined effects of read coverage and fragment length. The density of read ends in a cluster increases as read coverage increases. The uncertainty of the edge of the cluster is approximated as the average distance between read ends in the clusters.

The SPANNER program was also to discover tandem duplications. The primary difference with deletion detection was that supporting Illumina paired-ends with an orientation such that the reverse mapped end precedes the forward mapped end in genomic coordinates were selected. This corresponds to a “negative” paired-end span as measured in mapped genome coordinates. Tandem duplications were identified from paired-ends with a mapping distance at the low extreme (p-value < 0.04% for trio project and p-value < 0.02% for low-coverage project) of the corresponding library insert size distribution (which was defined allowing for negative mapping lengths between the read ends). The clustering logic was identical to deletions. Furthermore, equivalent selection criteria were used to remove potential mapping artifacts. Additional criteria were imposed on candidate tandem duplications based on proximity to annotated VNTRs and based on the local read coverage (read depth) for each sample. The read depth was measured by counting reads with the leftmost end falling into the predicted duplication (NREAD). A null-model estimate for NREAD was based on 100 random samples of the same event length and repeat content (EREAD). From this an effective copy number for the candidate duplication was calculated as copy-number = 2•NREAD/EREAD for each individual in the dataset. Furthermore, candidates with copy-number > 2.5•(1-NF/50), where NF is the number of supporting paired-ends for each individual were discarded. Lastly, candidates with duplication length > 250 bp with copy-number > 2.2 were discarded.

8.6. Genotyping of Structural Variants

We determined the allelic state of deletion polymorphism in individual genomes using Genome STRiP. Genome STRiP integrates diverse features of NGS data (paired-ends, read depth and split reads) using a Bayesian model to produce calibrated genotype likelihoods. These genotype likelihoods were then integrated with SNP haplotypes using the Beagle phasing/imputation algorithm to produce a final set of genotype calls and haplotype-informed genotype likelihoods.

Genotyping by Genome STRiP is described in detail in a companion manuscript. At a high level, Genome STRiP utilizes three classes of information about the allelic state of a deletion in each individual:
(1) read-depth-informative reads that map within the variation;
(2) informative paired-ends that map outside of the variation but are aberrantly spaced when mapped to the reference genome; and,
(3) informative split reads that map to the breakpoint junctions when the breakpoint locations and alternate allele sequence are known at base pair resolution.

Two sets of deletions were genotyped separately: Deletions discovered in the low-coverage samples and deletions discovered in the trios. In both cases, genotyping was performed in 156 samples using the low-coverage Illumina data after duplicate removal using Picard MarkDuplicates. Data from the six trio samples was included, downsampled to roughly 4x coverage. Data from all samples was pooled for genotyping.

Likelihoods from the three classes of informative read were combined in a Bayesian model to generate initial genotype likelihoods. These initial likelihoods were then integrated with SNP haplotype information using Beagle (v3.1) and a reference panel of SNP genotypes from HapMap 3 r^2, to yield posterior genotype likelihoods for each variant. The phasing step was performed separately in each population; trio parents and children were analyzed separately. After the incorporation of haplotype information into posterior genotype likelihoods using Beagle, sites with sufficient information for genotyping were selected using two filters: (1) minimum call rate of 50% across all three populations using a genotype quality threshold of 13 (95% confidence) and (2) Hardy-Weinberg equilibrium p-value > 0.01 in each of the three populations.

8.7. Validation of Structural Variants

8.7.1. PCR Validation of SVs

The rationale behind PCR validation experiments of SVs is that an insertion is expected to result in a larger DNA product (amplicon) compared to the reference allele, whereas a deletion is expected to result in a shorter amplicon. In PCR experiments, by examining an electrophoresis gel and comparing amplicon bands with molecular markers, deviations from the expected amplicon size when assuming presence of the reference allele can be scored. These deviations were used to examine the size of the respective deletion or insertion. Furthermore, amplicon patterns were used to discern homozygous from heterozygous SVs, using the rationale that both variant alleles can be amplified by the given PCR reaction in parallel, depending on the size of the SV.

8.7.1.1. Design of PCR Validation Experiments

Random locus selection: To enable calculating FDRs for independent SV callsets, we randomly picked 100 loci from each callset for subsequent PCR validation experiments. The randomization was carried out by randomly picking, without replacement, from the entire list of generated calls for each SV discovery callset.

Primer design: We implemented an iterative PCR primer design pipeline to ensure the specific placement of primers into unique regions falling into 150 bp windows flanking the inferred SV breakpoint region on either side. The primer3 algorithm (available from http://frodo.wi.mit.edu/primer3/) was used for primer placement, with the option “exclude primers matching onto known repeats”. In-silico PCR (available from http://genome.ucsc.edu/cgi-bin/hgPcr) was applied (with default parameters) with the primers designed by primer3 to test for the putative presence of alternative amplicons with similar, or smaller size. Primer pairs generating unique amplicons were kept and used in the PCR experiments. If primer pairs generated more than one amplicon at the given size (or at a smaller size), as judged by in-silico PCR, the primer positions were masked with ‘N’s and the primer design pipeline was re-initiated. If primer3 failed to identify suitable primers, the windows for primer design were iteratively increased by 150 bp on either side of the inferred SV. In ~25% of target regions, no primers could be designed within 2 kb of the inferred SV breakpoint regions; these cases were not tested in the PCR validation experiments.

Design of additional primer sets to validate insertions of transposable elements: In addition to the primer design used for all PCR verifications, a second primer set was designed for predicted retrotransposon insertion loci if, 1) no PCR product of a predicted amplicon size was obtained or 2) all subjects scored as absent for the insertion (potential false negative result due to poor PCR amplification of the larger fragment). These additional primers were designed within 500 bp of flanking sequence on either side of the predicted insertion site. The flanking sequence was masked for the presence of *Alu* elements. Each primer was tested with BLAT (Kent 2002) and in-silico-PCR was performed (http://genome.ucsc.edu) to ensure that the primer pair would amplify one unique PCR fragment. For SVA and L1 candidate insertions, internal primers were designed within the sequence of the respective human-specific retrotransposon consensus sequence. Since transposable elements cannot be validated reliably using array based (hybridization) methods, we carried out a relatively large number of PCR amplifications (>700) to examine transposable element polymorphisms in humans.

8.7.1.2. PCR Validation: Experimental Conditions

PCR experiments were carried out in four different laboratories, yielding similar success rates.

At WTSI, PCR primers were synthesized at Sigma. A control sample, NA15510 was used as a control for each variant. PCR was carried out with JumpStart REDAccuTaq LA DNA polymerase (Sigma-Aldrich) on a PTC-225 DNA Engine Tetrad Cycler (Bio-Rad) in 25ul reaction volumes. 15ng genomic DNA was used as template. The PCR protocol was as follows: Initial denaturation at 96°C for 30 sec; then 28 cycles of 94°C 5 sec, 58°C 30 sec, 68°C 8 min; and followed by an additional cycle of 68°C for 30 min. All PCR products were checked on 1% agarose gel for band visualization and scoring. In order to enable validation of small SVs – i.e., SVs at a size range (<100 bp) that impeded scoring the validations based on estimating amplicon size – the PCR products were purified with QIAquick PCR Purification Kit (QIAGEN #28104), and capillary sequenced with the PCR primer from either end.

At LSU, primers were obtained from Sigma. PCRs were carried out as follows: PCR amplifications were performed in 25 μl reactions in a 96-well format using either a Perkin Elmer GeneAmp 9700 or a BioRad i-cycler thermo-cycler. Each reaction contained 10-50 ng of template DNA; 200 nM of each oligonucleotide primer; 1.5 mM $MgCl_2$, 1X PCR buffer (50 mM KCl; 10 mM TrisHCl, pH 8.3); 0.2 mM dNTPs; and 1-2 U *Taq* DNA polymerase. For predicted amplicons larger than 2 kb LA-taq DNA polymerase (Takara Bio USA, Clontech Laboratories, Inc. Mountain View, CA) was used according to the manufacturer's instructions. For fragments up to 2 kb PCR reactions were performed under the following conditions: initial denaturation at 94°C for 90 sec, followed by 32 cycles of denaturation at 94°C for 20 sec, annealing at 57 ^{o} -61^{o} for 20 sec, and extension at 72°C for 30 to 90 sec depending on the predicted PCR amplicon size. PCRs were terminated with a final extension at 72°C for 3 min. For samples amplified with LA-taq DNA polymerase the same conditions were applied with the exception of the extension steps. Here, extension at 68°C was increased to 8:30 min and the final extension at 68°C was set to 10 min. For all SVA and some L1 insertion loci, a second (internal) PCR with one primer residing within the retrotransposon insertion was performed to verify insertion presence/absence. Primer sequences and PCR conditions can be found at http://batzerlab.lsu.edu. 20 μl of each PCR product were size-fractionated in a horizontal gel chamber on a 2% (*Alu* and SVA) or 1% (L1) agarose gel containing 0.1 μg/ml ethidium bromide for 60 minutes at 175V or 105min at 150V, respectively. Each agarose gel contained two lanes of DNA ladder, one 100 bp (cat.No. 170-8352) and one 500 bp (cat. No.170-8354) EZload^{TM} molecular ruler (BioRad Laboratories, Inc. Hercules, CA). UV-fluorescence was used to visualize the DNA fragments and images were saved using a BioRad ChemiDoc XRS imaging system (Hercules, CA).

The randomly selected retrotransposon insertion loci were analyzed with PCR on either a subset of 25 DNA samples from the low-coverage project or a panel containing the DNA of the two trios of the trio project. Both panels also included HeLa (ATCC CCL-2), a “Pop80” sample (a locally pooled DNA sample from different individuals of diverse geographic origins [Asia, Africa, South America, and Europe]) and a common chimpanzee sample (NS06006, Coriell). In addition, a population outgroup sample was included on the panel; in the case of the low-coverage project an individual of South American origin (NA17319, Coriell), and for the trio project a sample of Asian origin (NA17081, Coriell).

At Yale, PCRs were carried out using a previously described protocol (Korbel, Urban et al. 2007). In brief, PCR was carried out with JumpStart™ REDAccuTaq® LA DNA Polymerase (Sigma-Aldrich Inc., St. Louis, MO) on PTC-225 DNA Engine Tetrad™ Cycler (Bio-Rad, formerly MJ Research, Hercules, CA) in a 25 µl or 50 µl reaction volume and with 10 or 20 ng of genomic DNA as template. The following program was used: Initial denaturation at 94°C for 30 sec, followed by a 3-Step-Touchdown: 1. (94°C 5 sec, 68°C 30 sec, 68°C 6 min), 2. (94°C 5 sec, 66°C 30 sec, 68°C 6 min), 3. (94°C 5 sec, 64°C 30 sec, 68°C 6 min); followed by an additional cycle of 68°C 30 min.

At EMBL, PCRs were preformed using 10ng of NA12878 genomic DNA (Coriell) in 20 µl volumes in a C1000 thermocycler (BioRad). Two different enzymes, iProof High Fidelity DNA Polymerase (Biorad) and Hotstart Taq (Qiagen) were used, with comparable results. PCR conditions for iProof were : 98°C for 1min, followed by 5 cycles of 98°C for 10 s, 68°C for 20s and 72°C for 4 min and 30 cycles of 98°C for 10s, 66°C for 20s and 72°C for 4.5min, followed by a final cycle of 72°C for 5min. PCR conditions for HotStart Taq were: 94 °C for 15 min, followed by 5 cycles of 94°C for 30s, 60°C for 30s and 72°C for 3min and 30 cycles of 94°C for 30s, 56°C for 30s and 72°C for 3.5min, followed by a final cycle of 72°C for 5min. PCR products were analyzed on a 1% agarose gel stained with Sybr Safe Dye (Invitrogen) and a 100 bp ladder and 1 kb ladder (NEB).

8.7.1.3. Analysis of PCR Validation Data

Amplicons were analyzed in terms of size in comparison to molecular markers. For example, in order to validate retrotransposon insertion sites, amplicons matching the size predicted for the pre-insertion site were scored as a zero (0) while amplicons in agreement with the insertion were scored as a one (1). Individuals in which the insertion was homozygously absent were scored as ‘0,0’; if the sites was homozygously present we scored the insertion as ‘1,1’; and heterozygotes were scored as ‘1,0’.

8.7.2. Validation of SVs by Array Comparative Genome Hybridization (Array-CGH)

8.7.2.1. Microarray Design and Experimental Procedure

The validation of predicted structural variants via array-CGH was carried out in two stages. The first stage made use of custom Agilent 1M CGH microarrays to interrogate and validate/invalidate predicted deletions from the CEU trio individuals. To minimize bias, two independent array designs were created (Harvard and Agilent). A common design strategy was set as follows: a minimum of 5 oligonucleotide probes was used for each SV, and further 5 probes were included in the 5 kb flanking region of each SV. Uniquely mapped probes were used, where possible, though the repetitive nature of some structural variants necessitated the use of sub-optimal probe sequences.

The second stage of array-CGH validation made use of a high-resolution array set which was made available through a recent project from the Genome Structural Variation Consortium (Conrad, Pinto et al. 2010), whereby each trio individual was also analyzed with 20 custom 2.1M Nimblegen CGH microarrays. The design for this array set was such that oligonucleotide probes were placed approximately 50 bp apart across the entire human genome. Thus, this unbiased design would allow for the validation of many variants (known or novel) without the need for a design targeted to specific regions of the genome. These arrays were prepared using standard Agilent experimental procedures and those recently outlined (Conrad, Pinto et al. 2010), respectively.

8.7.2.2. Analysis of Agilent and Nimblegen Array-CGH Data

The nature of both array-CGH designs described above necessitated the development of custom analytical approaches to validated the predicted SV regions. Two complementary approaches were constructed, both of which utilized probes in regions of known copy number states (McCarroll, Kuruvilla et al. 2008) to build models of expected probe behavior.

The first method determined the extent the probes in each predicted structural variant region deviated from the expected null (i.e., from the homozygous reference allele, or the copy number ‘2’). This was done by building an empirical distribution of $\log_2$ ratio measurements in probes from regions with previously known ‘2:2’ (sample:reference) copy number states in both the samples and the reference and interrogating the probes from each region at both tails. A region observed as significantly deviating from this null distribution was deemed validated.

The second method utilized empirical distributions of $\log_2$ ratio measurements in regions of previously known ‘2:2’, ‘2:1’, and ‘1:2’ (sample:reference) copy number states to construct and arbitrate between null and alternative distributions using the relative likelihood of the observed data. Unlike the first approach, this method is not only able to *validate* but can also *invalidate* a particular region.

The sensitivity and specificity of both methods were determined through application and comparison using regions of known copy number state identified in a recent study (Conrad, Pinto et al. 2010). ROC curves were constructed for different structural variant size ranges to determine optimal thresholds for both methods.

8.7.2.3. SuperArray Validation

In addition to analyzing hybridized Agilent and Nimblegen arrays, SuperArray Validation (SAV) was used on array-based intensity data to validate deletion and duplication events and to estimate the false-discovery rate (FDR) of call sets. The "SuperArray" integrated available data from three array platforms (Affymetrix 6.0, Illumina 1M, and a custom Nimblegen aCGH array with 4,938,838 probes) into a high-density virtual array. Because array platforms differ in their quantitative response to underlying copy number, we developed a non-parametric test based on the simple assumption that, for any probe, samples with lower underlying copy number will, on average, tend to have a lower intensity measurement than samples with greater underlying copy number. Each structural variant call to be evaluated consisted of a genomic segment (chr, start, end) and a list of samples predicted to carry the putative deletion or duplication (copy-number different from '2'). Putative structural variants were evaluated by comparing the intensity measurements in duplications or deletions between the sample in which the event was predicted to the intensity measurements of all other samples, using the Wilcox rank sum test. For each probe falling into a putative structural variant, samples were ranked in intensity space, then the ranks of all probes for samples with inferred copy-number of '2' samples were compared to the ranks of all other samples. Rank data across all the probes within the putative deletion or duplication were then combined. For putative deletions, the expected intensity ranks for samples displaying a copy-number smaller than '2' are expected to be lower than for other samples; similarly, for putative duplications, the expected intensity ranks for samples displaying a copy-number larger than '2' are expected to be higher than for other samples. Putative deletions and duplications were considered validated if a significant Wilcox rank sum *P*-value of $P < 0.01$ was measured. The FDR for a call set was estimated as two times the fraction of putative calls for which we measured a Wilcox rank sum *P*-value of $P > 0.5$.

8.7.3. Validation of the Breakpoints of SVs by Array Capture

8.7.3.1. Microarray Design and Experimental Procedure

We attempted to validate 2,414 regions, for which deletions were predicted in NA12878, using a microarray-based sequence capture approach. For this purpose a custom Nimblegen microarray with probes covering 2 kb flanking regions of deletion breakpoints was designed at Yale. Array design was optimized to maximize the uniform coverage over target regions by using probes of ~75 bp in length containing unambiguously mappable sequence (i.e., the probe sequences have a single hit in build36/hg18 of the human reference genome). Overall, 65-82% of target regions were covered by probes. Genomic DNA from three samples corresponding to a parent offspring trio with European ancestry (daughter NA12878, mother NA12892, and father NA12891) was hybridized to the array. Captured DNA was sequenced using the 454 GS FLX Titanium platform, yielding approximately ~1x coverage per haplotype per sample.

8.7.3.2. Analysis of Array Capture Data

Reads were aligned to the human reference genome using Megablast and those mapped to the target regions were subsequently realigned using the Needleman-Wunsch algorithm with zero gap extension penalty (in order to allow for alignment extension across large gaps). Needleman-Wunsch alignments were post-processed by merging alignment fragments separated by less than 5 bp gaps and by removing fragments shorter than 20 bp. The breakpoints flanking the largest gap were compared to the predicted deletion breakpoints to validate the deletion. A deletion was considered as validated if: (1) the deletion and the largest gap displayed a reciprocal overlap of 50%, and (2) the sum of the discrepancies in breakpoint coordinate assignment was smaller than 5 kb (i.e., approximately twice the insert size used for 454 paired-end sequencing of NA12878).

8.7.4. Inference of the FDR and Construction of the SV Discovery Set

8.7.4.1. FDR Inference

The estimates of false discovery rate (FDR) for each algorithm showed generally a high concordance between PCR and array-based analysis, with the best concordances achieved on SVs discovered by algorithms estimating SV breakpoints at sufficient resolution for PCR and array-based validation approaches. The array- and PCR-based validation approaches had complementary strengths: since array-based copy number data was available on all samples (on 2-3 independent array platforms) and on a genome-wide scale, a very large number of putative SVs could be evaluated. Smaller SVs (<1 kb), which frequently did not have probes on array-based platforms, could generally be evaluated by PCR, but practical considerations limited the PCR validation to about 100 calls per call set. To integrate these FDR estimates into a single overall estimate of FDR for each algorithm, we calculated FDR hierarchically, using the array-based results to estimate FDR for all putative SVs that contained array probes, and extrapolating from the PCR-based FDR to estimate FDR for the remaining events in each call set. These overall FDR estimates for each call set are shown in Supplementary Tables 4A and 4B.

8.7.4.2. Selection of SV Calls for Constructing an Integrated Call Set

In creating an SV data release for the 1000 Genomes Project, we sought to realize the following two goals: (1) a global false-discovery rate less than 10%, such that more than 90% of SVs in the data release would correspond to real SVs; and (2) inclusion of the largest and most diverse set of SV calls possible. We therefore developed the following framework for identifying SV calls eligible for release:

(1) From the SV discovery algorithms that yielded call sets with an FDR less than 10% (Genome STRiP and SPANNER), we included all SV calls.
(2) From the SV discovery algorithms that yielded call sets with FDR greater than 10% (10 algorithms for low-coverage population data yielding FDR of 22-69%, and 14 algorithms for high-coverage trio data yielding FDR of 12-89%), we included the subset of SV calls that had been independently, explicitly validated by PCR or array-based experiments.

8.7.4.3. Integration and Merging of SV Calls Across Algorithms

To determine which of the SV discovery calls from each algorithm corresponded to the same, underlying SV, we first estimated a “confidence interval” for the bounds of each SV genomic segment identified by each algorithm. The size of these confidence intervals varied from algorithm to algorithm – extremely tight (single-base-pair) for algorithms based on split reads, fairly tight (mostly 2-40 bp) for algorithms based on paired-end mapping, and wider (about 1 kb) for algorithms based on read depth only. We “merged” SV calls for which the confidence intervals for both the left and right breakpoints overlapped, and estimated a revised confidence interval for the breakpoints of merged calls, from the overlap of the confidence intervals of all calls contributing to the merged call. 28,339 algorithm-level calls from the high-coverage trio data (across 14 algorithms) were merged into 11,321 independent SV discovery calls for the 1000 Genomes data release (Supplementary Table 4A); 34,085 algorithm-level calls from the low-coverage population data (across 10 algorithms) were merged into 15,947 independent SV discovery calls (Supplementary Table 4B).

8.8. Analysis of Breakpoint Junctions of SVs

We compiled a library containing the breakpoint junctions of approximately 14,000 SVs, i.e., all SVs from the project for which breakpoint coordinated have been mapped at single-nucleotide resolution (a full list is available on the project FTP site). The library was derived from seven individual SV call sets, which were partially derived through experiments (648 SVs) and through split-read or assembly-based algorithms (53,352 SVs). The experiments included PCR experiments and subsequent Sanger sequencing at WTSI, and array-capture experiments at Yale University. Split-read analysis was carried out with Pindel as well as with the Yale split-reads approach, whereas assemblies were generated with Cortex, TIGRA, and based on novel sequence insertions identified from orphan and one-end anchored inserts, respectively. In order to create a breakpoint junction library, the SVs were subjected to the following filtering steps:
- only SVs >=50 bp in size were retained;
- only validated calls were retained, i.e., we required either validation by PCR (based on amplicons size-scoring), array-CGH, array-capture or local breakpoint assembly;
- SVs with homology, or microhomology, at the breakpoint junctions were standardized to the left-most breakpoints;
- redundant SVs with the same chromosome, start- and end-coordinates were made non-redundant.

The breakpoint library was then analyzed by the BreakSeq pipeline (Lam, Mu et al. 2010) (downloadable at http://sv.gersteinlab.org/breakseq/). Using BreakSeq, SVs were classified according to their likely mechanism of formation, as previously described (Lam, Mu et al. 2010). In particular, SVs were classified into the following formation mechanisms: (1) Non-allelic homologous recombination (NAHR); (2) nonhomologous recombination (NHR), including nonhomologous end-joining (NHEJ) and fork stalling and template switching (FoSTeS/MMBIR); (3) variable number of tandem repeats (VNTRs); and (4) transposable element insertions (TEIs). Furthermore, the ancestral states of the SVs were inferred by aligning breakpoint junction sequences to the primate genomes as described previously (Lam, Mu et al. 2010). The results were converted to a standardized GFF format.

9. Genotype Accuracy as a Function of Depth in the Low-Coverage and Exon Projects

Forty-one CEU samples shared by both the exon and low-coverage projects were first identified. Both call sets were limited to the exon project target region (~1.43 Mb). Then, for each call, the genotype accuracy was computed using the following procedure: For every variant site, those non-reference SNP genotype calls that were greater and equal to the specific genotype depth threshold were tallied. Genotypes were compared with valid genotypes at HapMap II sites not in HapMap 3, and the total number of erroneous calls (i.e., non-variant genotypes according to HapMap) was counted. The procedure was repeated for various required genotype depth thresholds. Finally, the ROC curves of the required genotype depth thresholds for both projects were generated as shown in Figure 2d.

10. *De Novo* Assembly

A *de novo* genome assembly of the low-coverage project Illumina data (1.9Tb) was carried out with Cortex, a de Bruijn graph assembler (Caccamo, Iqbal et al. 2010). Each individual was first assembled separately, applying a base quality filter of 10, and removing PCR duplicates. A k-mer value of 29 was chosen which provides a good balance between sensitivity and disambiguation of repeat content. All individuals in a population (CEU, YRI or CHB+JPT) were then merged into a single graph. Finally an error-removal step was applied, of removing all k-mers that occurred only once within the population. The three populations were then loaded into a single multicolour graph that also contained the human reference autosome, and X and Y chromosomes (from build NCBI36) in three further colours. Novel sequence from this assembly was then determined by comparing contigs in the graph with the human reference genome. A contig was considered to be novel if every 29-mer within it was absent from the human reference genome. Contigs longer than 100 bp were blasted against the NCBI nucleotide and HuRef databases, and contaminants were removed (anything matching any non-primate species).

Given the short read-length (mostly 36 bp) and the choice of a relatively long k-mer (29 bp), this assembly is expected to have comparatively low polymorphism sensitivity. Of the 15 million SNPs called by the low-coverage project, we estimated the expected SNP sensitivity to be ~31% (Caccamo, Iqbal et al. 2010). In a graph with k = 29, 15 million SNPs would generate 15 million * 59 bp = 885 Mb of novel sequence in alternative alleles. Thus we expect to find 0.31*885 = 274 Mb. In fact a total of 261 Mb of novel sequence was found in SNP alternative alleles, close to expectations.

A total of 3.7 Mb of novel sequence longer than 100 bp were found, of which 87% matched known human or primate clone sequence in Genbank, and 79% matched the Venter (HuRef) genome. All of the contigs were validated by one or other of these two methods.

11. Imputation Analyses

11.1. Accuracy of Imputation Using 1000 Genomes Data as a Reference Panel

We investigated the performance of using the low-coverage project haplotype sets as reference panels for imputation. This was achieved by taking the HapMap 3 genotypes of the CEU and YRI trio fathers, and using these to impute the unobserved genotypes from the low-coverage project CEU and YRI haplotype sets respectively with IMPUTE (Howie, Donnelly et al. 2009). These unobserved genotypes were then compared to genotype calls from the trio project. The trio project genotypes are likely to be very accurate due to the high coverage sequencing used and so constitute a good benchmark dataset for comparing imputed genotypes. When imputing genotypes, it is much easier to impute homozygous genotypes for the common allele at each SNP, and so when comparing the imputed genotypes to the trio project genotypes we only considered those genotypes that had at least one copy of the minor allele. The minor allele was identified based on the allele frequencies in the low-coverage project haplotype sets and the results were stratified by minor allele frequency. The same analysis was repeated using the HapMap II CEU and YRI reference panels. The results are shown in Figure 4a.

11.2. eQTL Imputation

To evaluate the power of 1000 Genomes based analyses in genetic association studies, we re-analyzed two previous datasets (Dixon, Liang et al. 2007; Stranger, Nica et al. 2007). First, we used previously described methods (Stegle, Parts et al. 2010) to re-analyze microarray expression data available for a subset of the 1000 Genomes samples (Stranger, Nica et al. 2007). In this analysis, gene expression values were preprocessed using factor analysis to partially remove shared variation between transcripts, which can be due to experimental artifacts and features of the cell state. Then, we used the Mann-Whitney U test, which is robust to outliers, to test for correlation between adjusted transcript levels and genetic markers within 50 kb of the transcript. Empirical significance thresholds were derived through 1,000 permutations of the data. Second, we re-analyzed the microarray expression data of Dixon et al (Dixon, Liang et al. 2007). In this analysis, gene expression values were quantile normalized, to remove outliers, and then preprocessed using principal component analysis to remove shared variation between transcripts. The genotype data of Dixon et al (Dixon, Liang et al. 2007), which included genotypes from the Illumina 317K chip, was augmented using genotype imputation to also include markers present in the 1000 Genome Project CEU samples and which could be imputed with estimated $r^2 > 0.50$ (Li, Willer et al. 2009). Genotypes were imputed using MACH (Li, Willer et al. 2009) and the final association analysis was carried using the FASTASSOC procedure in Merlin (Chen and Abecasis 2007), which accounts for family correlations using a variance component model. As with the analysis of 1000 Genome samples, empirical thresholds were derived through 1,000 permutations of the data.

12. Evaluation of Mutation Rates

In this section, we outline our strategy for identifying and interpreting *de novo* mutations (DNMs) in the trios. Full details of the experiments and analyses can be found in a companion paper. Three groups (Broad Institute, Sanger Institute, and University of Montreal) generated a list of candidate DNMs from the trio project data using different methods (described below). These call sets were merged to create a single list of putative DNMs in each trio, and all calls in each merged set were taken forward to a validation stage. As the sequencing data are generated from lymphoblastoid cell lines (LCLs) and not primary cells, it was necessary to devise additional experiments to separate constitutional DNMs (that is, germline mutations present in either the egg or sperm that produced the trio child) from DNMs that are either somatic or cell-line in origin. We addressed this need during the validation stage, by sequencing additional samples that were informative about the germline status of each candidate DNM.

12.1. Discovery of *De Novo* Mutations with Trio Project Data

Broad. Broad Institute's discovery process was based on the same set of Broad SNP genotype calls that were used to create the release dataset for the trio project. Their genotyping process is described elsewhere in this supplementary material. A metric was created to summarize the support for a *de novo* mutation at each site by adding together three "lod scores". A lod score for each parent was defined as log10(L(homozygous reference | D)/ L(next best genotype | D)). The lod score for the child was log10(L(best fitting heterozygous genotype | D)/ L(homozygous reference | D)). This confidence metric was used to rank candidates for inclusion in the validation stage described below.

Sanger. The Sanger Institute group used a likelihood-based approach to assess the evidence of a DNM at each locus. This method provides the joint likelihood of the read-level data for all three trio members given a particular genotype configuration and the base at the reference sequence. A likelihood is assigned for all 1000 possible unordered, labeled genotype configurations that the trio may assume at a single site. This provides a natural way of accommodating triallelic SNPs. The method is Bayesian, as the calculations incorporate the prior probabilities of observing a *de novo* mutation, observing an inherited variant, and of the observed sampling configuration of derived alleles among parental chromosomes.

The ultimate output of this DNM caller is a posterior probability that a site contains a DNM, using the following approach. Let M, D, and C be 10-element vectors containing the likelihoods of all 10 possible genotypes given the mapping qualities and base qualities of the reads at the locus, for the mother, father, and child, respectively. In practice these likelihoods were generated by SAMtools 0.1.7 using the trio project BAM files created from Illumina data and released by the 1000 Genomes Project. Then a rescaled version of the joint trio likelihood surface is obtained with the following steps:

$P = M \otimes D$
$F = P \otimes C$
$T = F \circ R$
$X = T \circ Y$

Where '⊗' is the Kronecker product operation, 'o' is the Schur product operation, R is the matrix of transmission probabilities corresponding to each trio configuration, and Y is the matrix of priors corresponding to each trio configuration. The maximum likelihood trio configuration compatible with DNM, $x_{i\text{-max},j\text{-max}}$, is identified, and the posterior probability is calculated as:

$$\Pr(\text{de novo}) = \frac{x_{i-\max, j-\max}}{\sum_{i,j} x_{i,j}}$$

UdeM. The University of Montreal (UdeM) group developed a probabilistic method to identify candidate *de novo* mutations. The approach considers each genomic site separately and uses the aligned reads for each individual within a trio to simultaneously infer all three genotypes. The parameters of the UdeM model include: (1) the population mutation rate θ, which controls the expected heterozygosity of parental genotypes; (2) the germ-line mutation rate μ, which defines the rate at which the events of interest occur; (3) the somatic mutation rate μ_S, which models mutations arising anywhere between the germ-line and the cell line; and (4) the sequencing error rate ε, which quantifies the frequency with which a read differs from the sequenced genotype. The generative model is depicted in Supplementary Figure 16. Maternal and paternal genotypes are sampled from the population according to θ, and gametes are transmitted randomly and subject to mutation according to μ. Mutations in each individual that distinguish the sequenced genetic material (i.e., cell-line DNA) from its germ-line counterpart accumulate according to μ_S. Reads are modeled as random samples from the cell-line DNA, with error introduced according to ε. The model thus specifies $\Pr(R_M,R_F,R_O \mid \theta, \mu, \mu_S, \varepsilon)$, where the sequencing reads are observed (Supplementary Figure 16, shown as ovals) and the parameters can either be estimated by maximum likelihood (e.g., via expectation-maximization) or given a priori values. Bayes' Rule is used to make joint inference on the missing trio genotypes (Supplementary Figure 16, shown as rectangles), and the site-specific posterior probability that an offspring mutation (somatic or germ-line) has occurred can be calculated. (Note that in this trio design, the two mutational events cannot be disentangled.) Mapping and base quality scores thresholds were set at 20 and 10, respectively, based on the values reported within the trio project BAM files.

12.2. Filtering and Merging of Calls to Create Validation Lists

Our validation philosophy was to identify as many DNMs as possible, thus we used a permissive threshold for calling, generated a long list of variants for each trio, and attempted to validate all of these experimentally. After generating the candidate variant list, we anticipated that some filtering would be necessary; we know that our models do not capture all features of the data necessary to avoid convincing false positives. In order to assess the impact of our assumptions about what filters were necessary, we decided to leave the Broad set of candidates sites unfiltered in designing the study.

The Sanger and UdeM groups applied a common set of filters to their callsets. These filters fell into three broad categories: (1) proximity to other known variants, (2) overlap with primary genome sequence known to be problematic for mapping and assembly, and (3) other properties of the read-level data.

In each trio we included all post-filtered sites assigned a posterior probability greater than 10% by either the UdeM or Sanger method, as well as the top 500 unfiltered Broad calls not present in the union of the UdeM and Sanger sets. This led to 2750 candidate calls in the YRI trio and 3236 in the CEU trio.

12.3. Validation Experiments

We attempted to validate all 2750 candidate DNMs in YRI and all 3236 in CEU using two parallel approaches based on next-generation sequencing.

Samples. During validation, we sequenced genomic DNA from all 6 LCLs that were used to generate the trio data (but note, perhaps different lots). In order to separate germline from somatic (or cell-line) mutations, we screened additional DNA samples with the same validation assays. The CEU trio is part of a larger, 15 member CEPH/UTAH pedigree (1463, which includes the partner of NA12878 and 11 of her children). We included DNA from the 11 grandchildren to confirm germline status by inheritance. For the YRI trio (Y117), the Coriell Institute provided genomic DNA extracted from the same primary blood samples that were used to generate their LCLs.

Validation Experiment I. The Sanger Institute designed a pipeline to independently amplify DNM loci by nested PCR, pool these PCR products, and sequence the pools with a single lane of Illumina GA2 each. Pooled PCR products were sequenced separately for each DNA sample, except the 11 CEU grandchildren were pooled together post-PCR, and the PCRs from blood-derived DNA of the 2 YRI parents were pooled as well. Reads were aligned with BWA, sorted and indexed with SAMtools, and then post-processed with base quality recalibration and cleaning of small indels (via multiple sequence realignment) using the Broad Institute Genome Analysis Toolkit (GATK).

Validation Experiment II. UdeM resequenced all candidate *de novo* mutations using Agilent's SureSelect Target Enrichment System and ABI SOLID3 Plus sequencing. The

SOLiD3 Plus reads were aligned using ABI's Bioscope v1.2 software, and the resulting alignments were recalibrated using the GATK Base Recalibration tool.

Results. We created a unified analysis approach that jointly models the UdeM and Sanger validation data. The outcome of this analysis was to categorize each candidate DNM as one of the following: *de novo* germline mutation, *de novo* mutation only observed in the cell line, inherited variant present in the parents, and a false positive call (i.e., no variation observed in any sample). We require that the difference in log likelihood between the best fitting model and next-best fitting model be greater than some threshold, otherwise we consider the data uninformative and categorize the locus a "no call".

12.4. Calculation of Mutation Rate

The overall mutation rate provided in the main text is based on only for the portion of the genome analyzed by all 3 centres (i.e., not filtered by the UdeM/Sanger-specific filters).

This rate was calculated with the following equation:

Rate = [True Positives – False Positives + False Negatives] / Bases interrogated.

The total number of bases interrogated by all three algorithms was 2,555 Mb in CEU and 2,549 Mb in YRI, and in this fraction of the genome were 45/49 validated DNMs in CEU and 35/35 validated DNMs in YRI.

Based on our validation results the number of true positives in CEU is estimated to be

45 observed validated DNMs * [2802 attempted sites / 2197 called sites] = 57.39 DNMs

and for YRI

35 observed validated DNMs * [2332 attempted sites / 1782 called sites] = 45.80 NMs.

We estimated the experiment-wide false negative rate in two ways, by simulation and empirically (by comparing sensitivity of the different centres to the set of validated germline DNMs). Conservatively assuming complete dependence in the sensitivity of callers, both approaches suggest that we have missed 4% of the true DNMs in the portion of the CEU genome analyzed by all three groups, and 7% in the analogous section of the YRI genome. We believe that we have eliminated all false positives in the validation stage. These numbers then yield the following estimates of mutation rate:

CEU: 1.17×10^{-8} (95% CI: 0.94×10^{-8} - 1.73×10^{-8})
YRI: 0.97×10^{-8} (95% CI: 0.72×10^{-8} - 1.44×10^{-8})

13. Annotation of SNP Variants

Variants are annotated using the GENCODE gene models (Harrow, Denoeud et al. 2006) and the Human Genome Mutation Database (HGMD Professional, version 2009.4;Stenson, Mort et al. 2009). For the exon project, extrapolations to the whole genome are made assuming that the 1.4 Mb coding sequence in the target region represents a random sample from the total of 35.2 Mb in GENCODE.

13.1. Annotation of SNP Calls with Ancestral Allele

The human (NCBI36), chimpanzee (CHIMP2.1), orangutan (PPYG2) and rhesus macaque (MMUL_1) genomes were aligned using Enredo and Pecan (Paten, Herrero et al. 2008). In short, this method uses a set of conserved sequences to detect genomic point anchors (GPAs) in all four genomes. GPAs appearing too many times or in one genome only are filtered out. The resulting set is used to build the initial Enredo graph. Enredo proceeds through a series of graph transformations and simplification, resulting in a modified graph where the edges represent sets of co-linear segments. Duplications in one of the genomes appear as an edge where the same genome is represented two (or more) times.

Enredo's segments are then aligned with Pecan, a consistency-based multiple aligner optimized for genomic sequences. For segments not representing any duplication, Pecan can use the standard species tree. In the other cases, a sequence tree must be inferred. To do so, an initial quick alignment is built using a random tree and Pecan in pre-align mode. The resulting alignment is used to refine the tree, which in turn is used to refine the alignment. This process is run iteratively until convergence.

The ancestral alleles were derived from the Pecan alignments using Ortheus (Paten, Herrero et al. 2008). Ortheus uses a branch-transducer model - a type of Hidden Markov Model (HMM) - to infer insertion and deletion events. Substitutions are handled using a Tamura-Nei evolutionary model (Tamura and Nei 1993). Ortheus works progressively. During the initial phase, ancestral sequences are inferred from the extant sequences up to the root of the tree. However, uncertainties at this stage are modeled using sequence graphs, which allows Ortheus to defer choices until more information from other sequences is available.

Ancestral alleles are called using the immediate ancestor of the human sequence. In the vast majority of the alignments, this will be the human-chimpanzee ancestral sequence. In other cases, the ancestral sequence can correspond to the human-orangutan ancestral sequence or to other combinations, depending on the history of duplication and deletion events. A confidence level is assigned to each ancestral allele by analysing the neighbouring sequences in the tree, typically the chimpanzee sequence and the human-chimpanzee-orangutan ancestral sequence. A high confidence call is made when these other two bases match the ancestral one. If only one of the other two bases matches the ancestral allele, then the call is considered low confidence. In the cases where both disagree with the ancestral allele, the call is ignored.

13.2. HGMD Disease Variants Detected in the 1000 Genomes Data

The intersection between the 1000 Genomes data and HGMD Professional release 2009.4 missense and nonsense SNPs (Stenson, Mort et al. 2009) was identified for each of the projects separately based on the chromosome coordinates. Entries were retained when the HGMD disease allele was present in the 1000 Genomes calls and was tagged by HGMD as DM (damaging mutation, the most severe classification). We excluded sites where the disease allele in HGMD differs from both reference allele and alternative allele in the 1000 Genomes Project data. Note that although the disease allele is usually the derived allele, it is in some cases the ancestral allele. Intersections are listed for each project in Supplementary Table 5. The HGMD annotation in the column “disease” carries a question mark in a minority of entries; such entries were filtered out in some analyses.

In order to search for categories of disease that were under- or over-represented in the 1000 Genomes data, the following analyses were performed. The disease terms for 83% (83,648 mutations out of 100,023) of HGMD disease-causing mutations (Stenson, Mort et al. 2009) were mapped to disease concept identifiers (CUI) from the Unified Medical Language System (UMLS; 2010AA release; www.nlm.nih.gov/research/umls/), using a Java implementation of a simple word permutation-based method developed and tested by Shah et al. (Shah, Rubin et al. 2006; Shah and Muse 2008)

Using all disease concepts from HGMD as a control dataset (background), we compared the distribution of disease concepts in the HGMD - 1000 Genomes overlap subset (both heterozygous and homozygous 1000 Genomes calls) against the HGMD background dataset (Supplementary Figure 12) To allow for multiple testing, the significance of any difference noted was then assessed by means of Fisher’s Exact test with Bonferroni correction. Only p-values < 0.00278 (0.05/18, to allow for 18 tests) were considered significant.

13.3. Loss of Function Annotation

Functional annotation of SNPs, short indels and large structural variants was determined with reference to the GENCODE v3b annotation release (Harrow, Denoeud et al. 2006). Coding SNPs were mapped on to transcripts annotated as “protein_coding” and containing an annotated START codon, and classified as synonymous, non-synonymous, nonsense (stop codon-introducing), stop codon-disrupting or splice site-disrupting (canonical splice sites). Transcripts labeled as NMD (predicted to be subject to nonsense-mediated decay) were not used. Small deletions predicted to cause a frame-shift and large deletions predicted to disrupt gene function were also analysed.

Nonsense and splice-disrupting SNPs were flagged as likely representing reference error or annotation artefacts if the inferred loss-of-function (LOF) allele was also the ancestral state, or if the reference (non-LOF) allele was not observed in any individual in that population, and were excluded from the per-individual counts in Table 2. Splice-disrupting SNPs in non-canonical splice sites were also discarded. We did not consider the frame-shift status of splice-disrupting SNPs due to the challenges of inferring the effects of removal of splice-donor and acceptor sites on exon structure, but rather treated all such SNPs as likely to affect gene function.

We classified large deletions as gene-disrupting if they fulfilled the following criteria:

1. Removal of >50% of the coding sequence; or
2. Removal of the gene’s transcriptional start site or start codon; or
3. Removal of an odd number of internal splice sites; or
4. Removal of one or more internal coding exons that would be predicted to generate a frameshift.

For large deletions with imprecise breakpoints, we conservatively required that the deletions defined by both the inner and outer confidence intervals would have the same predicted effect on gene function. For cases with microhomology at the break-point we treated the breakpoint as falling to the right-hand side of the region of microhomology.

We did not perform functional annotation for large duplications due to the challenges of inferring functional consequences. The numbers stated in the text should thus be regarded as a lower bound for the number of observed loss-of-function variants per individual genome.

However, it should also be noted that the proportion of false positive calls in the LOF class due to sequencing and annotation errors is expected to be substantially higher than the genome-wide average. This effect is expected as LOF sites show a low level of true polymorphism due to selective constraint, meaning that a uniform error rate across the genome will result in a higher proportion of false positive calls compared to other (more variable) sites.

Enrichment of false positive calls in LOF variants is most evident in the CHB+JPT samples, which showed a higher per-individual number of LOF SNPs than other populations despite a comparable number of synonymous variants (Supplementary Table 11), as well as an unusual peak in the derived allele frequency spectrum (Supplementary Figure 13). This is likely due to a mild elevation in genome-wide false positive rates for SNPs in this population compared to other samples, which is then highly enriched at functionally constrained sites.

To lower the number of false positive indel calls we applied more stringent filters to the subset of indels that were called in the genome-wide set and were predicted to fall into the LOF class. The stringent filter requires that the range of positions where an indel would yield the same alternative haplotype sequence as the original called indel (for instance, in a repeat, the deletion of any repeat unit would give the same alternative haplotype), plus 4 bases of reference sequence on both sides of this region, was covered by at least one read on the forward strand, and at least one read on the reverse strand, with at most one mismatch between the read and the alternative haplotype sequence resulting from the indel (regardless of base-qualities). This filter removed an excess of 1 bp frameshift insertions seen in CHB+JPT with respect to CEU in the less stringently filtered genome-wide indel call set, although it is expected to remove a significant number of true positive calls as well. The indels that pass these stringent filters have been annotated in the project’s VCF files.

Experimental validation and manual reannotation of identified LOF variants is currently ongoing (manuscript in preparation).

For extrapolating the functional variants identified per individual in the exon project to the whole genome (Table 2) we used the ratio of the total coding sequence and splice sites in the exon-capture target regions (1,423,559 bp and 7,513, respectively) to the corresponding numbers for the GENCODE v3b annotation set as a whole (35,676,620 bp and 384,439, respectively).

The coordinates and predicted functional consequences of all of the LOF variants identified in the project are available on the 1000 Genomes FTP site.

14. Population Genetics

14.1. Detection of Selection

14.1.1. Identification of SNPs with Large Frequency Differences between Populations

The genotype calls for the low-coverage project were generated separately for each population. This makes comparison between populations challenging, as a SNP may be called only in one population, leaving an ambiguous frequency of the SNP in the other populations. Therefore, in order to accurately assess the evidence for frequency differentiation between the populations we used an alternative method that works directly from genotype likelihoods generated, as described above, by SAMtools. Specifically, using the dynamic programming algorithm outlined above that is used in the non-LD stage of the Sanger low-coverage genotype calls, we calculate (a) L_0, the maximum log likelihood of the data as a function of variant frequency assuming Hardy-Weinberg equilibrium across all samples from the analysis panels analysed and (b) L_1; the sum of the maximized log likelihoods for each analysis panel treated separately (and assuming Hardy-Weinberg equilibrium in each). The key feature of this approach is that it directly interrogates the evidence for differential allele frequencies between populations rather than relying on call sets. We use a Likelihood Ratio Test (LRT) to identify sites with strong evidence of frequency differences between the populations. Under the null hypothesis of no differentiation $2(L_1-L_0)$ should be approximately chi-squared distributed with degrees of freedom equal to the number of analysis panels considered minus one. To reduce the effects of differential mapping we only considered genotype likelihoods computed from Illumina data. Estimates of the allele frequency difference between populations were also obtained by this approach and are presented in Supplementary Tables 7 and 8.

We performed the LRT on each of the pairwise population comparisons and selected all sites with a LRT statistic > 30. This provided a set of filtered SNPs with strong evidence of differentiation between the populations (Figure 5c). We selected sites with an absolute frequency difference greater than 0.8, and refer to these SNPs as highly differentiated. In total, there are 5,660 such SNPs in the CEU vs YRI comparison, 861 SNPs in the CEU vs CHB+JPT comparison, and 14,401 SNPs in the CHB+JPT vs YRI comparison.

14.1.2. Composite Likelihood Selection Statistic

To localize signals of selection, we implemented a method that combines multiple tests for selection, the Composite of Multiple Signals (CMS; Grossman, Shylakhter et al. 2010). Five statistics were used as inputs to composite: 1) F_{ST}, 2) ΔDAF, and three metrics of haplotype length, 3) iHS, 4) ΔiHH, and 5) XP-EHH (Voight, Kudaravalli et al. 2006; Sabeti, Varilly et al. 2007; Grossman, Shylakhter et al. 2010). Statistics were computed as described previously with the following modification for full sequence data. EHH was defined as the probability two randomly chosen chromosomes carrying the core allele are identical at polymorphisms greater than or equal to 5% for the interval from the core to point x (instead of all polymorphisms for the entire interval from the core to the point x) (Sabeti, Reich et al. 2002). Ancestral information was taken from the project's 4-way EPO alignments.

The CMS values in the plots represent the within-region localization scores, the probability that each SNP is the causal SNP conditional on being within 1 Mb of a selective event. These scores are optimal for fine-mapping signals of selection, but do not necessarily represent the strength of the selective event. We assess the strength of selection at each locus by calculating the probability each SNP is selected, without assuming that it is near a selected variant. We then calculate empirical p-values by comparing the value at each SNP to the background genomic distribution. At the EDAR locus shown in the main text, the within region localisation scores for the Asian population and CEU are similar, however the genome-wide strength of the selective signals are quite different. The highest scoring SNPs in the Asian populations have empirical p-values around 10^{-6}, among the highest in the genome, suggesting strong selection at the EDAR locus in the Asian populations. In the CEU population, the most extreme p-values are around 10^{-4}, possibly suggesting weaker selection.

To further analyze whether the pattern of scores at the EDAR locus is consistent with selection upon the V370A mutation, we used *cosi* to generate 1200 coalescent simulations of the 1 Mb regions with the recombination map from the EDAR locus. We used a calibrated demographic model of European, East Asian, and West African populations (Schaffner, Foo et al. 2005). Each replica contained a single positively selected allele at the position of the V370A mutation in the Asian population. To characterize the most distribution of scores resulting from selection at this position with this recombination map, we recorded the position of the highest scoring variant by CMS in each replica.

14.1.3. Neutrality Tests Based on Allele Frequency Spectra

The summary statistics, Tajima's D (Tajima 1989) and Fay and Wu's H (Fay and Wu 2000) reflecting aspects of the allele frequency spectrum, were calculated using a custom Java script for each non-overlapping ~10 kb window across the genome. The start and end chromosomal coordinates of each window are the first and last SNP positions in that window. For the few positions where the ancestral state of a segregating site was unknown, the major allele in YRI was used; this is a conservative assumption for the H test. For both tests, p-values were calculated using *ms* incorporating the best-fit demographic model (Hudson 2002; Schaffner, Foo et al. 2005), conditioning on the number of the segregating sites in each window, as well as empirical p values from the genomic distribution.

Two composite likelihood-ratio (CLR) analyses were performed on the same ~10 kb windows as the above summary statistics. The CLR analysis of Kim and Stephan (Kim and Stephan 2002) uses information contained in the frequency spectrum to calculate the likelihood ratios of a selection model versus the neutral model. The more significant the local reduction of genetic variation is, the higher the CLR value will be. An average recombination rate of 10^{-8} per base per generation was used, and local mutation rates were estimated by the program itself. The other CLR analysis, from Nielsen and colleagues (Nielsen, Williamson et al. 2005), is similar to the Kim and Stephan test but differs in that a neutral population model is not used; instead, a null distribution is derived from the general pattern of variation in the data itself. In our application, the frequency spectrum of each chromosome was used as the background spectrum in the analysis. Again, a high CLR value indicates a candidate for selection. For both tests, empirical p-values were calculated based on whole genome data.

14.2. Diversity in Genic Regions

14.2.1. SNP Diversity and Divergence in Genic Regions

The plot of average diversity levels across genic annotations (Figure 5a – upper panel) was generated by collating the longest protein-coding transcript for each autosomal protein-coding gene in GENCODE v3b with annotated start and stop codons. Genes with only a single exon are included in the "1st Exon" annotation. Genes with two exons are included in the "1st Exon" and "Last Exon" annotations, with the intron included with "1st Intron". Genes with at least three exons are included in all annotations. Note that "Other Introns" corresponds to a single intron being chosen from each gene (excluding the first intron). Diversity levels at a given distance were calculated as the average heterozygosity (2pq) across genes. Within an annotation, there are 200 points, corresponding to diversity in the first and last 25 base pairs, and with the remaining 150 positions sampled with uniform spacing across the element. Elements shorter than 150 base pairs were not considered. The “middle exon” corresponds to a single exon in the middle of the transcript, and “other introns” to a single intron chosen at random (but not including the first intron). The red curve was obtained by loess, with a smoothing parameter of 0.7 and polynomial degree 2. Plots for the two other population samples look very similar.

For the plot of average diversity levels divided by average divergence for different genic annotations (Figure 5a – lower panel) the average diversity at a given position was divided by the average divergence between human and chimpanzee. Divergence was calculated based on comparing hg18 to panTro2 using alignments downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/). Other details are as above. As can be seen, the diversity levels divided by divergence are remarkably similar across annotation. Additional details and interpretation are given in (Hernandez, Kelley et al. 2010).

14.3. Recombination Analyses

Low-coverage project genotype calls were used to estimate recombination rates from patterns of linkage disequilibrium. We used the interval program of the LDhat package (McVean, Myers et al. 2004; Auton and McVean 2007), which implements a Bayesian reversible-jump Markov chain Monte Carlo scheme to fit a piecewise-constant model of recombination rate variation. To estimate recombination rates on the autosomes, we first split the data in to sections consisting of 4000 SNPs, with a 200 SNP overlap between sections. On each section, we ran the program *interval* for a total of 30,300,000 iterations, taking a sample from the Markov chain every 15,000 iterations. The first 20 samples (corresponding to 300,000 iterations) were discarded. The process was repeated for CEU, CHB+JPT and YRI separately.

The samples consist of an estimate of the population recombination $\rho=4N_e r$ / kb between every consecutive pair of SNPs, where N_e is the effective population size, and r is the per-generation recombination rate. Large gaps in the genome, such as the centromeres, were set to have a recombination rate of zero recombination, as no reliable estimate for these regions could be obtained.

To convert our estimates into per-generation recombination rates, we need an estimate of the effective population size. This was achieved by comparison to the deCODE genetic map (Kong, Gudbjartsson et al. 2002), which provides an estimate of the per-generation recombination rate at the broad scale. We performed a linear regression (without intercept) of the deCODE rate estimate against the LDhat estimates at the 5 Mb scale, the gradient of which provides an estimate of $4N_e$. The Pearson correlation between the LDhat and deCODE estimates at this scale is approximately 0.90 for all populations (0.9151 CEU, 0.9059 CHB+JPT, 0.9150 YRI). The resulting estimates for N_e for each of the populations were: 11,603 for CEU, 13,002 for CHB+JPT and 20,163 for YRI.

To determine the increased resolution to detect hotspot boundaries, we used the hotspots detected from HapMap II (2007). For each hotspot, we identified the peak rate within the hotspot boundaries from the 1000G rate estimates. New boundaries for these hotspots were determined by identifying the point at which the estimated recombination rate dropped below 50% of the peak rate. This procedure is similar to that used to originally determine the hotspot boundaries in HapMap (Simon Myers; personal communication). Hotspots for which no peak could be detected were excluded. Of the 32,990 hotspots from HapMap II, a peak in the 1000G rates could be identified in 32,062 cases. The mean width of hotspots was found to be 2,336 bp (95% C.I. 2316-2358 bp from 1000 bootstrap samples), compared to 5,505 bp (95% C.I. 5467-5543 bp from 1000 bootstrap samples) for the original HapMap hotspots.

A potential positive correlation between meiotic recombination and diversity has been suggested by a number of previous studies (Nachman 2001; Lercher and Hurst 2002; Hellmann, Ebersberger et al. 2003; Spencer, Deloukas et al. 2006; Hellmann, Mang et al. 2008). To investigate whether such an effect exists, we developed an approach robust to obvious potential biases introduced by the fact that recombination patterns are themselves typically inferred using patterns of LD at segregating sites in the genome. To achieve this, we examined diversity levels in the CEU and YRI populations, as measured by the number of identified SNPs per base, surrounding occurrences of a previously identified human hotspot motif (Myers, Freeman et al. 2008; Myers, Freeman et al. 2008), CCTCCCTNNCCAC. This approach removes the requirement to identify recombinationally active loci using variation patterns directly. Because positions at which this motif occurs have roughly a 3-fold increased average recombination rate relative to the average for the genome, any strong effect on diversity levels caused by recombination hotspots would be expected to manifest as unusual levels of diversity surrounding motif sites.

We firstly obtained all genome-wide occurrences of the “hotspot motif” CCTCCCTNNCCAC in the human genome. To allow for potential sequencing biases caused by the genomic environment in which this motif is found, we also identified all positions of a control, “non-hotspot” motif C**T**TCCCTNNCCAC, differing by only one base from the hotspot motif, but showing no evidence in the low-coverage project data of an elevated recombination rate at motif sites (Figure 8c). In each case, we only considered occurrences outside masked repeats (identified using the RepeatMasker track downloaded from the UCSC database). The non-hotspot motif was chosen to have a similar frequency in the genome to the hotspot motif (3401 and 2735 occurrences respectively).

We identified all SNPs that fall into non-overlapping 100 bp windows spanning 10 kb to the left and right of motifs. To account for differences in, for example, sequence uniqueness near motif sites, we used the “callability” score which for each base in the genome gives a 0-1 measure defining whether a SNP passing QC filters (based on average sequencing depth, and non-zero mapping score) could be identified at that base. Average callability profiles around the hotspot and non-hotspot motifs were very similar. We measure diversity in a population within a given bin as the proportion of callable bases at which a SNP passing filters was found. Confidence intervals for the SNPs per base within each bin were calculated based on 1.96 standard errors of the mean diversity within the bin. Average recombination rates were estimated as described above in the 10 kb each side of occurrences of the motif and binned in the same way as for the diversity levels. The YRI diversity levels and recombination rates were used to construct Figure 8c, although the patterns observed in other populations are very similar.

14.4. Y chromosome Haplogroups

A maximum likelihood haplogroup tree under a HKY model of evolution was produced using phyML, and bootstrap values were produced using 100 subsamplings. Trees were produced using both all 2870 filtered sites (Supplementary Figure 7), and the 1971 UYR sites; though there was very little difference between the two trees. The haplogroup tree classifies all the major haplogroups as monomorphic, and recovers the relationships between them, with high bootstrap confidence. It also shows evidence for a deep division between haplogroups DE and CT, previously identified only by a single marker (P143; Karafet, Mendez et al. 2008). New insights into recent human evolution can also be gained from the branch lengths; for example, the short internal branch lengths within the haplogroup R1b relative to the other haplogroups suggest a recent expansion of this European haplogroup (Balaresque, Bowden et al. 2010).

15. Full Project Expectations

15.1. Increasing Read Lengths

We simulated paired-end reads from the human reference genome build 36 with 1% uniform sequencing error rate. The standard deviation of the insert size distribution is 10% of the average insert size. The simulated reads were then mapped to the genome with bwa-0.5.8 (Li and Durbin 2010). A read is considered to be mapped confidently if bwa assigns a mapping quality no less than 10. The left hand side of Supplementary Figure 15 shows the percent mapping confidently as a function of read length. For paired-end reads, the average insert size is fixed at 400 bp. It is evident that increasing read length improves the accessibility of the genome, especially for single-end reads. The right hand side of Supplementary Figure 15 shows the percent mapping confidently as a function of insert size with read length fixed at 100 bp. Increasing insert size also helps the accessibility, but not as much as increasing read length.

15.2. Predicting the Rate of Variant Discovery in the Full Project

To estimate the rate at which variants of different frequencies have been identified in the low-coverage project and to make predictions about the rate of discovery in the full project we use a simple, statistical model of population differentiation (Balding and Nichols 1995). In our model an ancestral population gives rise to a large (infinite) number of daughter populations, each related to each other to the same degree, described by a drift parameter, c. The allele frequency of a variant in a given subpopulation is described by a beta distribution with parameters x/c and (1-x)/c, where x is the frequency of the variant in the ancestral population. The average variant frequency across daughter populations is therefore x and the variance of the variant frequency across populations, divided by x(1-x) is c/(1+c), leading to the equation of c/(1+c) with Wright's Fst. To obtain the joint distribution of variant frequencies across populations conditioning on a given frequency in a given population we assume that variants are distributed with density proportional to 1/x in the ancestral population and use MCMC to estimate the conditional density distribution of variants in the ancestral population.

To predict the rate at which variants of different frequencies are discovered we simulate the number of variants present in the sampled individuals by modeling the sample counts in each population through the binomial model with parameters n_i and p_i, where n_i is the number of haploid genomes sampled from population i and p_i is the frequency of the variant in population i (simulated from the ancestral allele frequency and the beta model described above). The distribution of sample counts is combined with empirical estimates of power at different allele counts from comparison of the low-coverage project data to HapMap II genotypes in overlapping individuals (Figure 2a and Supplementary Figure 8). The contribution of discovery from populations in a different geographical region is not considered (neither are local migration and the spatial relatedness of populations).

16. Production Centre Protocols

Where methods differ from standard protocols, detailed methods are given below.

16.1. Baylor College of Medicine

16.1.1. Whole Genome 454 Sequencing

A mix of fragment libraries and paired-end libraries were constructed for each of the assigned samples. GS FLX fragment libraries and GS FLX Titanium fragment libraries were generated using 5ug and 10ug of genomic DNA respectively following standard methods (Genome Sequencer FLX and Genome Sequencer FLX Titanium Methods Manual). For this process the DNA was fragmented by nebulization to an average size of 500 or 700 bp, end repaired and specific adaptors added by ligation followed by purification and strand selection.

Long Paired-end libraries were constructed using 5ug of genomic DNA following standard methods (GS FLX Paired End DNA Library Preparation Method Manual). Briefly, fragments of 2-3 kb were produced by HydroShearing and biotinylated hairpin adaptors added that when cleaved with EcoRI, provide ligatable, cohesive ends for circularization. The circularized fragments were nebulized to a few hundred base pairs, end repaired and immobilized using the biotinylated linker for the addition of platform specific adaptors.

Each fragment and paired-end library was run on the Agilent Bioanalyzer 2100 to determine the library size, and the concentration determined by Ribo/PicoGreen assays. Libraries were then sequenced on the 454 FLX/Titanium platform using standard vendor emPCR, enrichment and sequencing methods.

16.1.2. Whole Genome SOLiD Sequencing

A combination of fragment and mate-paired libraries were utilized for sequencing on the SOLiD System V2.0 platform. SOLiD fragment libraries were constructed with 30ug of input DNA and an average insert size of 160 bp following standard vendor protocols (SOLiD System 2.0 Fragment Library Preparation) and sequenced as unidirectional reads of 25 bp. DNA input ranged from 30 to 50ug for mate-paired libraries to insure library complexity (SOLiD System 2.0 Mate-Paired Library Preparation). Insert sizes of 1.5 and 2.5 kb were utilized for sequencing in the 2 X 25 bp format providing both sequence coverage and structural information. Each fragment and paired-end library was run on the Agilent Bioanalyzer 2100 to determine the library size, and the concentration determined by PicoGreen assays. Each fragment and mate-paired library was then sequenced on the SOLiD V2.0 platform following standard vendor methods for emPCR, enrichment, 3' end modification and ligation sequencing (Applied Biosystems SOLiD System 2.0 User Guide).

16.1.3. Exon Capture and Sequencing

NimbleGen 385K capture chips were designed to target the 1000 gene regions. Pre-Capture libraries were constructed using 20 µg of genomic DNA following standard NimbleGen protocols (NimbleGen Sequence Capture: Short Library Construction Protocol and NimbleGen Arrays User's Guide). For this process the DNA was fragmented by nebulization to an average size of 700 bp and then subjected to end repair. NimbleGen linkers (gsel3 and gsel4) were ligated and then amplified using LM-PCR methods. Each sample was hybridized to a NimbleGen 385K capture chip and eluted using 95^{o}C H2O and then amplified again using LM-PCR methods. Quantitative PCR assays using SYBR Green were performed on a standardized set of 4 control loci present on all the arrays as a quality control measure. Final 454 platform sequencing libraries were constructed using 5ug of the amplified captured material using standard vendor protocols (GS FLX General Library Preparation Method). Capture libraries were then sequenced on the 454 FLX/Titanium platform using standard vendor emPCR, enrichment and sequencing methods (GS FLX Titanium Sample Preparation Manual).

16.2. Beijing Genomics Institute

16.2.1. Whole Genome Illumina Sequencing

Genomic DNA-Seq Pair-End libraries were generated from 3-5 ug genomic DNA using the Paired-End Genomic Sample Prep Kit (Illumina), as the manufacturer's instructions. Purified genomic DNA was sonicated and ligated to Illumina Pair-End DNA adapters (Illumina), after gel purification the adapter ligated DNA molecules around 500 bp (± 20 bp) were enriched by 10 cycles of PCR with primers complementary to the adaptor sequences. The concentration of the DNA library is measured by qPCR by Sybrgreen (ABI) on StepOne (ABI) and the size distribution is measured by Agilent 2100 bio-analyzer.

The purified Genomic DNAs were sheared, polished and prepared using the Illumina Index Sample Preparation Kit (Illumina). DNA libraries were amplified independently using 15-cycles of PCR amplification with PCR index primers. Amplified libraries were again size selected using agarose electrophoresis. After spin column extraction and quantitation, libraries were mixed at equimolar ratios to yield a multiplexing library.

Cluster generation on the Cluster Station (Illumina) is a process by which a denatured DNA fragment (the template) is hybridized to the surface of a specially grafted flow cell and amplified (bridge PCR) to form a surface bound colony (the cluster). Millions of different DNA fragments can be generated simultaneously to seed the surface of a single flowcell, leading to a heterogeneous cluster population, each cluster consisting of many identical copies of the original template molecule. It consists of a number of sequential sub-routines designed to seed, grow the clusters. We sequenced the cluster generated flow cell on GA/GA II/GA IIx following the manufacture's protocol.

After genome analyzer sequencing, Illumina GAPipeline was used for data analysis to get short reads (fastq), which includes 3 steps, image analysis, base calling and short reads alignment. In this pipeline, the image analysis module is Firecrest and the base caller is Bustard. Finally, Eland was used to map reads onto the human genome.

Based on alignment, we counted several quality criteria for quality control, which includes DNA libraries and sequencing bases. For DNA libraries, we compute the duplication rate and insert size duplication. We also calculated error rate, GC content, mapped rate and the base quality.

16.2.2. Whole Genome SOLiD Sequencing

For SOLiD Long Mate-Pair Library sequencing, 25 ug of DNA was sheared by HydroShear (Genomic Solutions) and end repaired with the Endit kit (Epicentre). Then the sheared DNA was ligated with EcoP15I cap adaptors size-selected to an average size of 1500 bp. The EcoP15I cap adaptors were left dephosphorylated so that circularization of the target DNA left a nick on the 3' ends of the internal adaptor. These nicks were bi-directionally extended into the insert DNA using a timed nick translation reaction. Tags were liberated with S1 nuclease, end repaired with the Epicentre Endit kit and varied in size from 50-75 bp per tag. All libraries were ligated by P1 and P2 adaptors (Applied Biosystems SOLiD™ Mate-Paired Library Oligos Plus Kit #4425772) with T4 DNA ligase (New England Biolabs) and amplified for 10 cycles (Applied Biosystems SOLiD™ 3 System Long Mate Pair Library Protocol).

We sequenced the mate-paired libraries with the Applied Biosystems SOLiD™3.0 System analyzer according to the manufacturers' instructions. For data analysis, we used SOLiD^{TM} System Analysis Pipeline Tools (Corona Lite) to map the SOLiD reads (csfasta) onto the human genome. Based on the alignment, we calculated the mapped rate, error rate, base quality and GC content for base quality control and computed the distribution of insert size and duplication rate for DNA libraries quality control.

16.3. Broad Institute

16.3.1. Whole Genome Illumina Sequencing

Genomic DNA was sheared to a range of 100-700 bp, end repaired and ligated to Illumina paired end adaptors. Fragments were purified and size selected (380 bp +/- 10%) using gel electrophoresis. Excised fragments were enriched over 10 cycles of PCR, denatured and normalized prior to cluster amplification and sequencing.

Cluster amplification of denatured templates occurred according to manufacturer's protocol (Illumina). Flowcells were sequenced on the Genome Analyzer II Instrument, using Sequencing-by-Synthesis kits and analyzed with the Illumina extraction pipeline. Standard quality control metrics including error rates, % passing filter (PF) reads, and total Gb produced were used to characterize process performance prior to downstream analysis. Each whole genome library was sequenced over several flowcells until an average genome coverage of 4x (low-coverage project) or 12x (trio project) in passed filter (PF) bases was achieved.

16.3.2. Exon Capture and Sequencing

Method I: Direct Sequencing of Hybrid Selected Pond Fragments
DNA oligonucleotides, corresponding to 170 bp of target sequence (1000 genes) flanked by 15 bp of universal primer sequence, were synthesized in parallel on an Agilent microarray, then cleaved from the array. The oligonucleotides were PCR amplified, then transcribed *in vitro* in the presence of biotinylated UTP to generate single-stranded RNA "bait." Genomic DNA was sheared, ligated to Illumina sequencing adapters, and selected for lengths between 200-350 bp. This "pond" of DNA was hybridized with an excess of bait in solution. The "catch" was pulled down by magnetic beads coated with streptavidin, then eluted. Hybrid selected libraries were denatured and normalized prior to cluster amplification and sequencing. Each library was sequenced on 1 - 2 lanes of an Illumina GA-II sequencer, using 76 bp paired-end reads.

Method II: Sequencing of Concatenated and Sheared Hybrid Selection Libraries
DNA oligonucleotides, corresponding to 170 bp of target sequence (1000 genes) flanked by 15 bp of universal primer sequence, were synthesized in parallel on an Agilent microarray, then cleaved from the array. The oligonucleotides were PCR amplified, then transcribed *in vitro* in the presence of biotinylated UTP to generate single-stranded RNA "bait." Genomic DNA was sheared to a range of 50-700 bp and ligated to Illumina sequencing adapters (GFA). This "pond" of DNA was hybridized with an excess of bait in solution. The "catch" was pulled down by magnetic beads coated with streptavidin and then eluted. Resulting enriched fragment "catch" was digested with Not1 enzyme and ligated to create concatenated fragments of > 1 kb which were subsequently sheared to 150 bp. Sheared and selected "catch" fragments were end repaired and ligated to Illumina sequencing adapters and enriched over 12 PCR cycles. Hybrid selected libraries were denatured and normalized prior to cluster amplification and sequencing. Each library was sequenced in 1-3 lanes of an Illumina GA-II sequencer, using 36 bp fragment reads.

16.4. Illumina

16.4.1. Whole Genome Illumina Sequencing

Preparation of short-insert paired-end Illumina sequencing libraries, flow cell preparation and cluster generation have been described previously (Bentley, Balasubramanian et al. 2008). Briefly, genomic DNA samples (5 µg) were randomly fragmented by nebulisation and used to prepare paired-end sequencing libraries with average insert sizes of 220 bp. Libraries were denatured using NaOH (0.1 N) and diluted in cold (4 °C) hybridisation buffer (5x SSC + 0.05 % Tween 20) to a working concentration of ~6 pM, prior to seeding clusters on the surface of the flow cell at a density of ~360,000 clusters per mm^2. Cluster amplification, linearization, blocking and hybridisation to the Read 1 sequencing primer were carried out on a Cluster Station using the Illumina Cluster Generation kit v1. Following the first sequencing read, flow cells were held *in situ* and clusters were prepared for Read2 sequencing using the Illumina Paired-End Module with the Cluster Generation kit v1. Paired-end sequence reads of 50 bases were generated using the Genome Analyzer II with v2 SBS reagent kits, as described in the Illumina Genome Analyzer operating manual.

Image analysis, base calling and quality scoring were carried out using the Illumina analysis pipeline version 1.3. Sequence reads were filtered out from those clusters whose proximity to others resulted in mixed sequence data (purity-filtering). On average 10.6 Gb of purity-filtered sequence data were generated per sequencing run using the described configuration of chemistry, instrumentation and analysis pipeline, and averaged across all runs 82.5% of base calls were estimated to have an accuracy of at least 99.9% (Q30). Sequence reads were aligned to the human NCBI36.1 reference sequence using ELAND to provide quality control information about the sequencing run.

16.5. Life Technologies

16.5.1. Whole Genome SOLiD Sequencing

Samples were sequenced using a combination of mate-paired libraries and fragment libraries with the Applied Biosystems SOLiD™ System (Life Technologies, Carlsbad, CA) according to the manufacturers' instructions. Mate-pair libraries were prepared with the TypeIII restriction endonuclease EcoP15I (Smith, Malek et al. 2004; Applied Biosystems SOLiD Library Preparation Guide). Additionally, sheared "fragment" libraries were generated and sequenced as unidirectional reads. Briefly, fragment libraries were generated by shearing genomic DNA to a 60-90 bp range using various shearing methods (DNaseI, Nebulization, and adaptive focused acoustic bombardment with a Covaris S2) and end repairing the DNA (McKernan, Peckham et al. 2009).

Emulsion PCR was performed according to Dressman et al (Dressman, Yan et al. 2003) with a few minor modifications (Applied Biosystems SOLiD Library Preparation Guide). Since limited dilution of DNA is utilized to produce clonal bead amplification, 70-80% of the beads in any given emulsion are un-amplified beads. An enrichment step is performed to select for the templated beads and provide a higher number of sequence generating features per run. Enrichment of amplified beads was performed as previously described (Shendure, Porreca et al. 2005) with a few modifications. Once emulsions are broken the beads are enriched, end modified and deposited on a microscope slide ready for SOLiD sequencing (Applied Biosystems SOLiD Library Preparation Guide) (McKernan, Peckham et al. 2009).

Ligation sequencing was performed in five different frames of sequencing as instructed by the manufacturer. As a result five different 5'phosphorylated primers that are each offset by 1 base with respect to each other are used. The detection probes have a cleavable phosphorothiolate linkage fixed between the 5th and 6th base such that sequencing with 1 primer generates partial dinucleotide information in 5 base increments. Primer 1 will survey dinucleotides 1,2 and 6,7 and 11,12 and so on to bases 46 and 47. Primer 2 will survey dinucleotides 0,1 and 5,6 and 10, 11, … 45, 46. Primers 3, 4 and 5 will be nested more than 2 bases into the known adaptor sequence and thus do not require their 1st ligation cycle to imaged (McKernan, Peckham et al. 2009).

16.6. Max Planck Institute for Molecular Genetics

16.6.1. Whole Genome Illumina Sequencing

Pilot project Illumina data were generated using the Genome Analyzer II (GAII, Illumina). Libraries were prepared from genomic DNA fragmented by ultrasound. 185-235 bp DNA fragments were gel purified and further processed into GAII paired-end (PE) libraries.

Libraries were prepared using Illumina PE library preparation kit. Several modifications were introduced in the original Illumina library preparation protocol (e.g., additional gel-purification after library amplification, which helps to get rid of unspecific PCR products; real-time check of non-amplified libraries for determination of required number of amplification cycles and estimation of library complexity; real-time check of 10nM library stocks before loading them onto flowcell to reach optimal cluster density) to make the process more reproducible and predictable. Libraries were loaded onto PE sequencing flowcells (the average cluster density was ~12x10^4 per tile). 36 bp PE runs were performed for each flowcell, allowing recognition of 36 nucleotides from each side of the genomic DNA insert.

Raw data were pipelined according to corresponding manufacturer's instructions. Base calling was performed using Illumina's Genome Analyzer Sequencing Control Software (SCS). Resulting sequencing reads were aligned to the human genome (hg18, NCBI build 36.1). For each sample, HapMap genotype validation was performed.

16.6.2. Whole Genome SOLiD Sequencing

For the SOLiD sequencing platform (version 2), fragment libraries (50 – 100 bp) were prepared using the ABI protocol with several modifications. End-repair reaction was performed according to the Illumina protocol. For test amplification and large scale amplification the Invitrogen mix was replaced by the 2x Phusion HF Master Mix (NEB, #F-531L). Resulting beads with attached library molecules were loaded onto the flowcell (amount of usable beads varied from 200 to 300 mlns per single-frame flowcell). For each flowcell, 35 bp fragment run was performed.

Raw data were pipelined according to corresponding manufacturer's instructions. Base calling was performed using SOLiD Analysis Tools. Resulting sequencing reads were aligned to the human genome (hg18, NCBI build 36.1). For each sample, HapMap genotype validation was performed.

16.7. Roche

16.7.1. Whole Genome 454 Sequencing

Genomic DNA for each sample was obtained from Coriell. Random shotgun libraries were generated by fragmentation of 6 mg human genomic DNA using the GS FLX Titanium General Library Preparation Kit following manufacturer's recommendations (454 Life Sciences, A Roche Company, Branford, CT, USA). Briefly, DNA was randomly sheared via nebulization and double stranded DNA adaptors were blunt ligated to fragment ends following post-electrophoresis agarose gel excision of the 500-800 bp fraction. The final single stranded DNA library was isolated via streptavidin bead binding to biotinylated adaptors followed by alkaline treatment. The library was then quantitated via fluorometry using Quant-iT RiboGreen reagent (Invitrogen, Carlsbad, CA, USA) prior to emulsion PCR amplification.

Genomic shotgun library molecules were clonally amplified via emulsion PCR following manufacturer's recommendations employing the GS FLX Titanium LV emPCR Kit (454 Life Sciences). Following amplification, emPCR reactions were collected and emulsions broken according to manufacturer's protocols. Beads containing sufficient copies of clonally amplified library fragments were selected via the LV enrichment procedure and counted with a Multisizer 3 Coulter Counter (Beckman Coulter, Fullerton, CA) prior to sequencing.

Following emulsion PCR enrichment, beads were deposited into the wells of a Titanium Series PicoTiterPlate device and 454 Sequencing was performed using the GS FLX instrument according to manufacturer's recommendations (454 Life Sciences). All sequencing employed the 2-region gasket format with 2 million enriched beads loaded per region. GS FLX Titanium Sequencing Kit XLR70 reagents were employed in all sequencing runs. Image analysis, signal processing and base calling were performed using a 2.0 pre-release version of the GS FLX Titanium Data Processing Software.

16.8. Sanger Centre

16.8.1. Whole Genome Illumina Sequence

Libraries were prepared and sequenced essentially as described in elsewhere (Bentley, Balasubramanian et al. 2008). 5 µg of genomic DNA from each sample was fragmented using a disposable nebulizer (Invitrogen) and purified using a qiaquick column (Qiagen). DNA was end-repaired as described (Bentley, Balasubramanian et al. 2008) and an adaptor ligated to the ends of the DNA (adaptor sequences: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATCxT (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG). Fragments of approximately 200 bp were gel-purified and PCR amplified. Flow cells were prepared, clusters generated, and processed flowcells were paired-end sequenced with 36-37 cycles each end on an Illumina Genome Analyzer and data processed using standard methods.

16.8.2. Exon Sequence Capture

20 ug of DNA were sheared to 100 – 400 bp using a Covaris S2 following manufacturer's protocols and the settings Duty Cycle, 20%; Intensity, 5.0; Cycles / burst, 200; Duration, 90; Mode, Freq Sweeping. Sheared samples were quantitated on a Bioanalyzer (Agilent, Santa Clara, USA). 10 – 15 ug of sheared DNA were end-repaired, A-tailed and Illumina sequencing adapters ligated to the resulting fragments using the Illumina Paired-End DNA Sample Prep protocol with the slight modification that the gel size selection step was replaced with a SPRI bead purification (following manufacturer's protocol). 5 ug of each library were hybridised to a custom Nimblegen 385-K array following manufacturer's protocols (Roche/Nimblegen) with the modification that no pre-hybridisation PCR was performed. Captured samples were washed and eluted in 50 ul of PCR-Grade water following manufacturer's protocols. Eluted samples were amplified using a master-mix containing 2 mM $MgCl_{2,}$ 0.2 mM dNTPs, 0.5 uM PE.1. 0.5 uM PE.2 and 3 units of Platinum® Pfx DNA Polymerase per sample. Samples were aliquoted into 3 individual wells of a plate and amplified using the following conditions: 94°C for 5 minutes followed by 20 cycles of 94°C for 15 seconds, 58°C for 30 seconds, 72°C for 30 seconds and a final extension of 72°C for 5 minutes. PCR products were purified using SPRI beads prior to sequencing.

Captured libraries were sequenced on the Illumina GA platform as paired-end 37-bp reads.

16.9. Washington University in St. Louis

16.9.1. Whole Genome Sequencing

Illumina fragment libraries were prepared using 1 ug of high molecular weight genomic DNA according to manufacturer's instructions, and the resulting libraries were sequenced on Illumina GA sequencers to produce approximately 2X coverage in fragment end reads of 36-50 bp for each genome.

16.9.2. Exon Capture and Sequencing

Biotinylated capture probes were generated using a synthetic probe library constructed to selectively target the 1,000 genes. The probe set was combined with each whole genome shotgun Illumina library, hybridized, and the resulting probe:library hybrids isolated using streptavidin magnetic beads. The captured library fragments were reclaimed by denaturation and sequenced as fragment end reads on the Illumina GA sequencer.

Capture oligonucleotides 190 bp in length (150 internal bases flanked by 20 bp PCR primers) were designed to tile end-to-end across target region. The resulting pool of synthetic oligos was amplified by PCR and incorporated with biotin-14-dCTP to produce a biotinylated capturing library (BCL). Genomic DNA was nebulized into 200-500 bp fragments, end-repaired, and then ligated with Illumina paired-end adapters to create a target library (TL) for each sample. Each target library was hybridized with the 500-gene BCL and then amplified by PCR. The resulting capture fragments were sequenced on the Illumina GAIIx platform. Sequence data were converted to FastQ format and mapped to the Hs36 reference sequence using MAQ v0.7.1.

17. References

Abyzov, A., A. E. Urban, et al. (2010). "CNVnator: An Approach to Characterize and Genotype Atypical CNVs Using High-throughput Sequencing Coupled with Population and Family Structure." Genome Res **Submitted**.

Albers, C. A., G. Lunter, et al. (2010). "Dindel: Accurate indel calls from short read data." in prep.

Alkan, C., J. M. Kidd, et al. (2009). "Personalized copy number and segmental duplication maps using next-generation sequencing." Nat Genet **41**(10): 1061-1067.

Andrews, R. M., I. Kubacka, et al. (1999). "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA." Nat Genet **23**(2): 147.

Auton, A. and G. McVean (2007). "Recombination rate estimation in the presence of hotspots." Genome Res **17**(8): 1219-1227.

Balaresque, P., G. R. Bowden, et al. (2010). "A predominantly neolithic origin for European paternal lineages." PLoS Biol **8**(1): e1000285.

Balding, D. J. and R. A. Nichols (1995). "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity." Genetica **96**(1-2): 3-12.

Bandelt, H. J., P. Lahermo, et al. (2001). "Detecting errors in mtDNA data by phylogenetic analysis." Int J Legal Med **115**(2): 64-69.

Bandelt, H. J., A. Salas, et al. (2004). "Artificial recombination in forensic mtDNA population databases." Int J Legal Med **118**(5): 267-273.

Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Brandon, M. C., E. Ruiz-Pesini, et al. (2009). "MITOMASTER: a bioinformatics tool for the analysis of mitochondrial DNA sequences." Hum Mutat **30**(1): 1-6.

Browning, B. L. and Z. Yu (2009). "Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies." Am J Hum Genet **85**(6): 847-861.

Caccamo, M., Z. Iqbal, et al. (2010). "Cortex – A unified framework for variant calling and genomic assembly." in prep.

Chen, K., J. W. Wallis, et al. (2009). "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." Nat Methods **6**(9): 677-681.

Chen, W. M. and G. R. Abecasis (2007). "Family-based association tests for genomewide association scans." Am J Hum Genet **81**(5): 913-926.

Comaniciu, D. and P. Meer (2002). "Mean shift: A robust approach toward feature space analysis." Ieee Transactions on Pattern Analysis and Machine Intelligence **24**(5): 603-619.

Conrad, D. F., D. Pinto, et al. (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-712.

Demichelis, F., S. R. Setlur, et al. (2009). "Distinct genomic aberrations associated with ERG rearranged prostate cancer." Genes Chromosomes Cancer **48**(4): 366-380.

DePristo et al (2010). "A framework for variation discovery and genotyping using next-generation DNA Sequencing Data for medical and population genetics projects." in prep.

DePristo, M., E. Banks, et al. (2010). "A framework for variation discovery and genotyping using next-generation DNA Sequencing Data for medical and population genetics projects." **Submitted**.

Dixon, A. L., L. Liang, et al. (2007). "A genome-wide association study of global gene expression." Nat Genet **39**(10): 1202-1207.

Dressman, D., H. Yan, et al. (2003). "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations." Proc Natl Acad Sci U S A **100**(15): 8817-8822.

Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**(3): 1405-1413.

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.

Hajirasouliha, I., F. Hormozdiari, et al. (2010). "Detection and characterization of novel sequence insertions using paired-end next-generation sequencing." Bioinformatics **26**: 1277-1283.

Harris, R. S. (2007). Improved pairwise alignment of genomic DNA, Pennsylvania State University.

Harrow, J., F. Denoeud, et al. (2006). "GENCODE: producing a reference annotation for ENCODE." Genome Biol **7 Suppl 1**: S4 1-9.

Hellmann, I., I. Ebersberger, et al. (2003). "A neutral explanation for the correlation of diversity with recombination rates in humans." Am J Hum Genet **72**(6): 1527-1535.

Hellmann, I., Y. Mang, et al. (2008). "Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals." Genome Res **18**(7): 1020-1029.

Hernandez, R., J. L. Kelley, et al. (2010). "Classic selective sweeps were rare in recent human evolution." in prep.

Hormozdiari, F., C. Alkan, et al. (2009). "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes." Genome Res **19**(7): 1270-1278.

Howie, B. N., P. Donnelly, et al. (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.

Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." Bioinformatics **18**(2): 337-338.

Irwin, J. A., J. L. Saunier, et al. (2009). "Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples." J Mol Evol **68**(5): 516-527.

Karafet, T. M., F. L. Mendez, et al. (2008). "New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree." Genome Res **18**(5): 830-838.

Kent, W. J. (2002). "BLAT-the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Kidd, J. M., G. M. Cooper, et al. (2008). "Mapping and sequencing of structural variation from eight human genomes." Nature **453**(7191): 56-64.

Kim, Y. and W. Stephan (2002). "Detecting a local signature of genetic hitchhiking along a recombining chromosome." Genetics **160**(2): 765-777.

Knuth, D. E. (1968). The art of computer programming. Reading, MA, Addison-Wesley Pub. Co.

Kong, A., D. F. Gudbjartsson, et al. (2002). "A high-resolution recombination map of the human genome." Nat Genet **31**(3): 241-247.

Korbel, J. O., A. Abyzov, et al. (2009). "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data." Genome Biol **10**(2): R23.

Korbel, J. O., A. E. Urban, et al. (2007). "Paired-end mapping reveals extensive structural variation in the human genome." Science **318**(5849): 420-426.

Lam, H. Y., X. J. Mu, et al. (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." Nat Biotechnol **28**(1): 47-55.

Le, S. Q. and R. Durbin (2010). "SNP detection and genotyping from low coverage sequencing data on multiple diploid samples." in prep.

Lercher, M. J. and L. D. Hurst (2002). "Human SNP variability and mutation rate are higher in regions of high recombination." Trends Genet **18**(7): 337-340.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics **26**(5): 589-595.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res **18**(11): 1851-1858.

Li, N. and M. Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." Genetics **165**(4): 2213-2233.

Li, R., H. Zhu, et al. (2010). "De novo assembly of human genomes with massively parallel short read sequencing." Genome Res **20**(2): 265-272.

Li, Y., C. Willer, et al. (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.

Li, Y., C. J. Willer, et al. (2010). "MaCH: Using sequence and genotype data to estimate haplotypes." Genet. Epi. **32**: 1-19.

Lunter, G. and M. Goodson (2010). "Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads." in prep.

Marchini, J., B. Howie, et al. (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." Nat Genet **39**(7): 906-913.

Marth, G. (2010). "Mosaik-aligner." from http://code.google.com/p/mosaik-aligner/.

Marth, G. T., I. Korf, et al. (1999). "A general approach to single-nucleotide polymorphism discovery." Nat Genet **23**(4): 452-456.

McCarroll, S. A., F. G. Kuruvilla, et al. (2008). "Integrated detection and population-genetic analysis of SNPs and copy number variation." Nat Genet **40**(10): 1166-1174.

McKenna, A., M. Hanna, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

McKenna et al (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Submitted.

McKernan, K. J., H. E. Peckham, et al. (2009). "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding." Genome Res **19**(9): 1527-1541.

McVean, G. A., S. R. Myers, et al. (2004). "The fine-scale structure of recombination rate variation in the human genome." Science **304**(5670): 581-584.

Minichiello, M. J. and R. Durbin (2006). "Mapping trait loci by use of inferred ancestral recombination graphs." Am J Hum Genet **79**(5): 910-922.

Myers, S., C. Freeman, et al. (2008). "A common sequence motif associated with recombination hot spots and genome instability in humans." Nat Genet **40**: 1124-1129.

Myers, S., C. Freeman, et al. (2008). "A common sequence motif associated with recombination hot spots and genome instability in humans." Nat Genet.

Nachman, M. W. (2001). "Single nucleotide polymorphisms and recombination rate in humans." Trends Genet **17**(9): 481-485.

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

Paten, B., J. Herrero, et al. (2008). "Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs." Genome Res **18**(11): 1814-1828.

Paten, B., J. Herrero, et al. (2008). "Genome-wide nucleotide-level mammalian ancestor reconstruction." Genome Res **18**(11): 1829-1843.

Sabeti, P. C., D. E. Reich, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.

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

Schaffner, S. F., C. Foo, et al. (2005). "Calibrating a coalescent simulation of human genome sequence variation." Genome Res **15**(11): 1576-1583.

Schwartz, S., W. J. Kent, et al. (2003). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-107.

Shah, N. H. and M. A. Muse (2008). "UMLS-Query: a perl module for querying the UMLS." AMIA Annu Symp Proc: 652-656.

Shah, N. H., D. L. Rubin, et al. (2006). "Ontology-based annotation and query of tissue microarray data." AMIA Annu Symp Proc: 709-713.

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Simpson, J. T., K. Wong, et al. (2009). "ABySS: a parallel assembler for short read sequence data." Genome Res **19**(6): 1117-1123.

Smith, D. R., J. A. Malek, et al. (2004). Methods for producing a paired tag from a nucleic acid sequence and methods of use thereof, Google Patents.

Spencer, C. C., P. Deloukas, et al. (2006). "The influence of recombination on human genetic diversity." PLoS Genet **2**(9): e148.

Stegle, O., L. Parts, et al. (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." PLoS Comput Biol **6**(5): e1000770.

Stenson, P. D., M. Mort, et al. (2009). "The Human Gene Mutation Database: 2008 update." Genome Med **1**(1): 13.

Stranger, B. E., A. C. Nica, et al. (2007). "Population genomics of human gene expression." Nat Genet **39**(10): 1217-1224.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.

Tamura, K. and M. Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." Mol Biol Evol **10**(3): 512-526.

The International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **In press**.

The International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

van Oven, M. and M. Kayser (2009). "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation." Hum Mutat **30**(2): E386-394.

Voight, B. F., S. Kudaravalli, et al. (2006). "A map of recent positive selection in the human genome." PLoS Biol **4**(3): e72.

Wang, L. Y., A. Abyzov, et al. (2009). "MSB: a mean-shift-based approach for the analysis of structural variation in the genome." Genome Res **19**(1): 106-117.

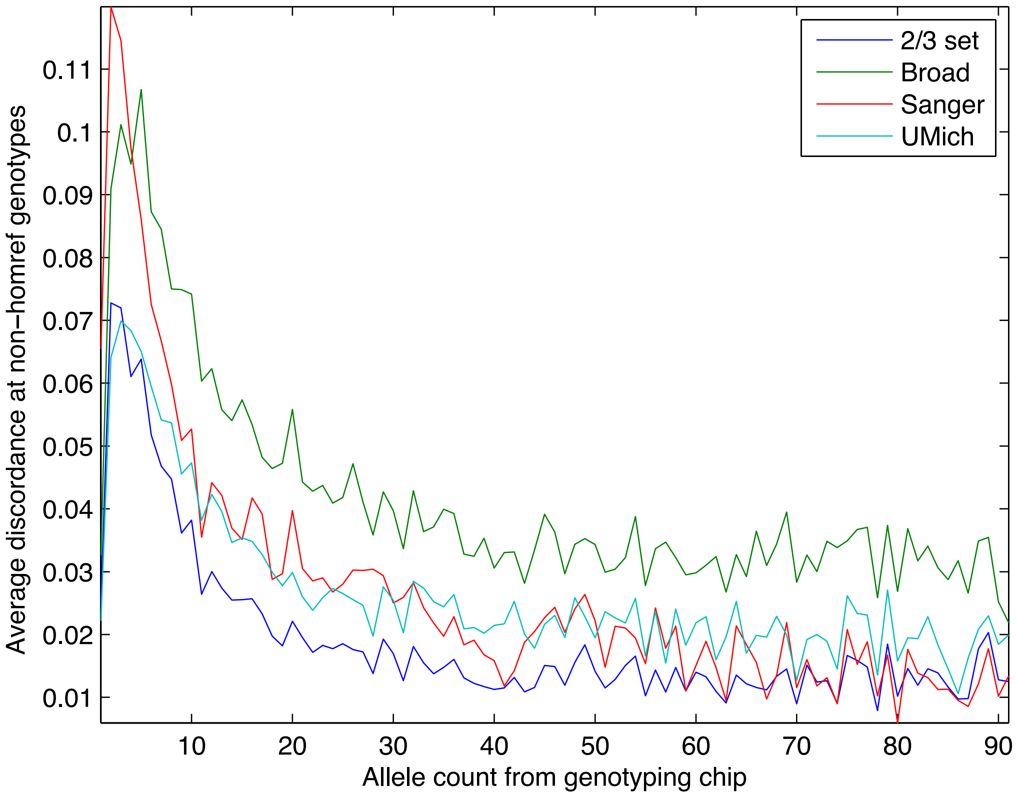
Wheeler, D. A., M. Srinivasan, et al. (2008). "The complete genome of an individual by massively parallel DNA sequencing." Nature **452**(7189): 872-876.

Ye, K., M. H. Schulz, et al. (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." Bioinformatics **25**(21): 2865-2871.

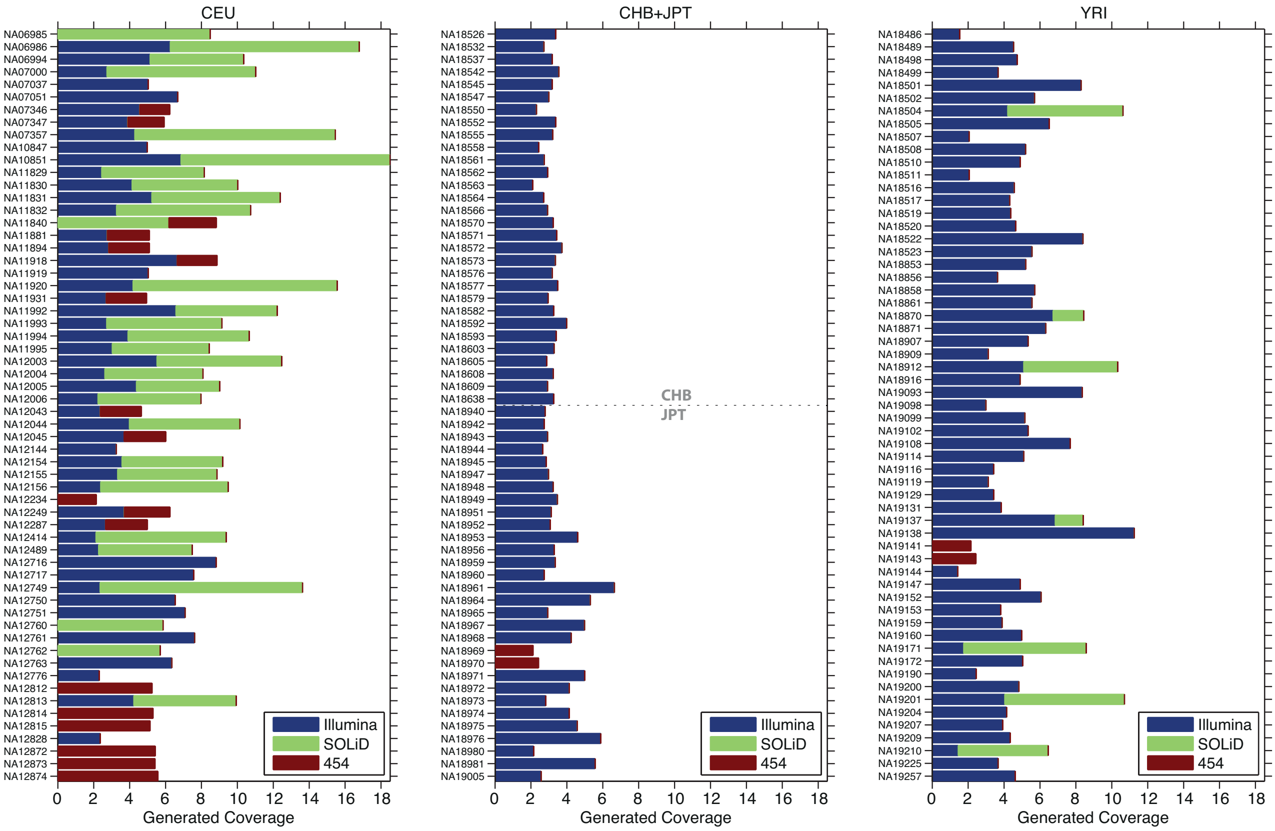
Yoon, S., Z. Xuan, et al. (2009). "Sensitive and accurate detection of copy number variants using read depth of coverage." Genome Res **19**(9): 1586-1592.

Youssef, S. (1987). "Clustering with local equivalence relations." Computer Physics Communications **45**: 423-426.

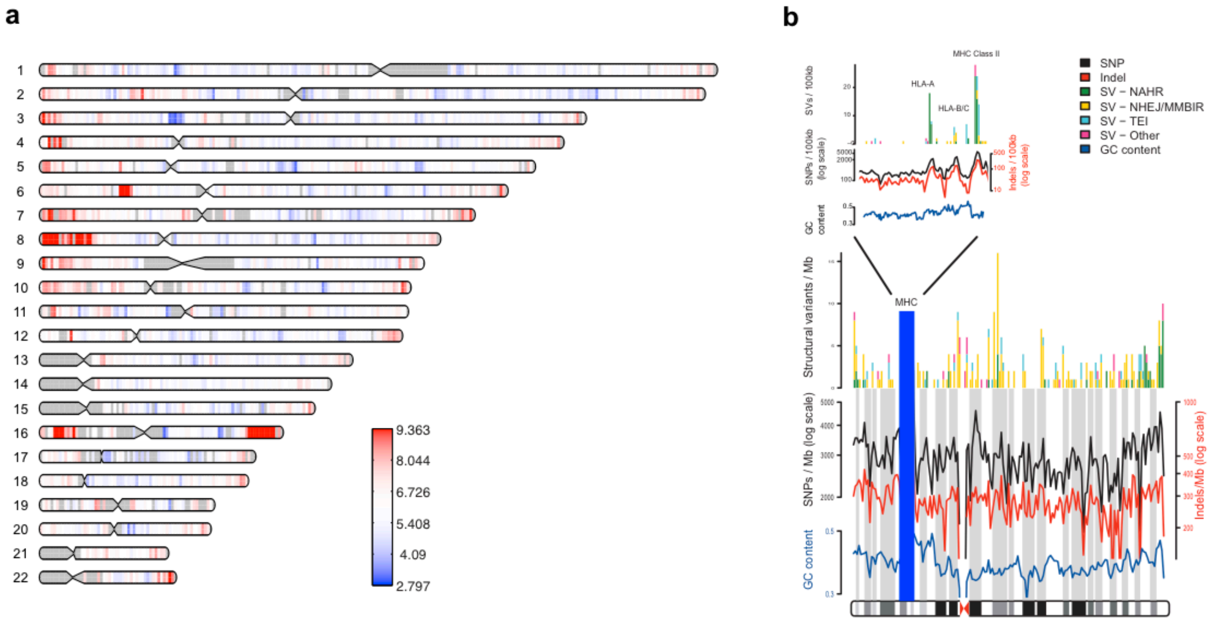
18. Supplementary Figures



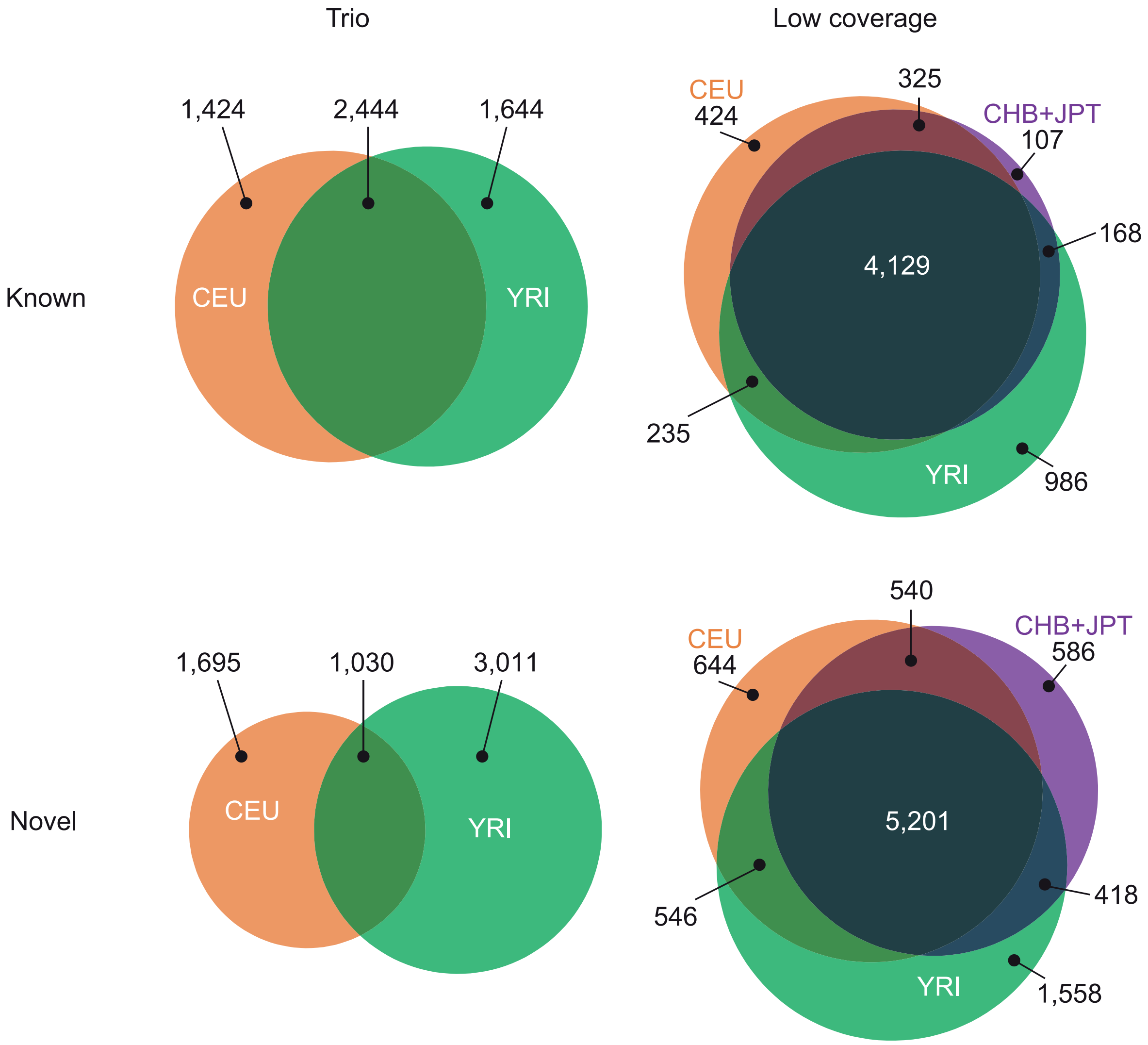
Supplementary Figure 1. Average discordance between primary and consensus low coverage genotype calling methods and 1000G genotyping chip (see Section 6.6). Discordance at sites called on the chip as variant (i.e., not homozygous for the reference allele) was calculated in 46 overlapping CEU samples at 50,367 sites that were called by all three primary methods (Sanger, Michigan and Broad) and that were also polymorphic according to the genotyping chip. The consensus genotypes (2/3 set) have consistently lower discordance than any single call set. Overall, the consensus set makes between 25% and 50% fewer errors than the individual call sets.



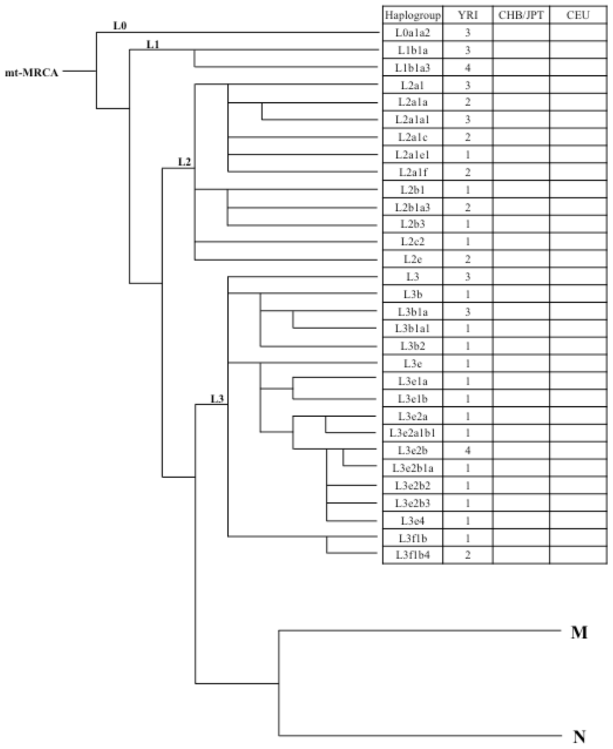
Supplementary Figure 2. Amount of sequence coverage generated (mapped bases/2.85 Gb) in the low-coverage project by sample and sequencing technology; blue = Illumina, green = SOLiD, red = 454. Note that populations and samples differ considerably in coverage (CEU highest, CHB+JPT lowest, sample coverage from c. 2x to 18x) and the balance of technologies. Many samples have data from two technologies.



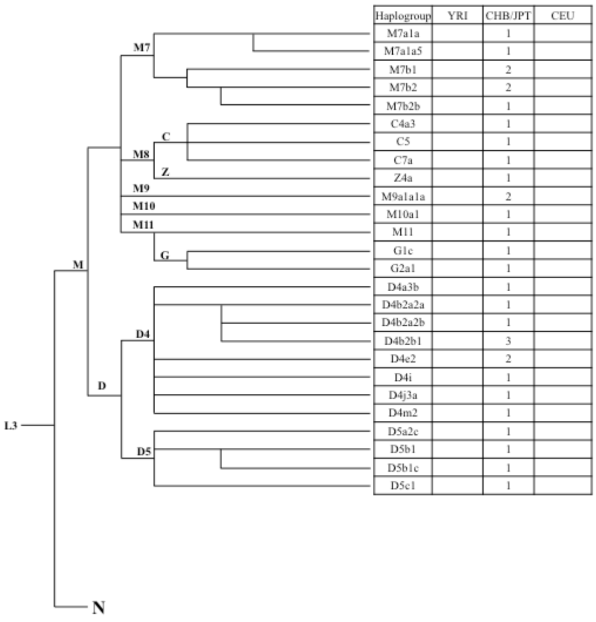
Supplementary Figure 3. Variation across the genome. **a,** Genomic distribution of SNP density in the low-coverage analysis on the autosomes. The colours show the SNP density (SNPs / kb) in 1 Mb bins, with red indicating higher densities and blue indicating lower densities (Kin and Ono 2007). SNP densities 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 5 Mb region of very low diversity on 3p21. The regions of high SNP density on 8p and chromosome 16 coincide with regions of extensive structural variation. **b,** Distribution of variants on chromosome 6 in the low-coverage project. From the bottom upwards are shown GC content (blue), the density of small indels (red) and SNPs (black) in the CEU samples, 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 HLA 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 HLA.



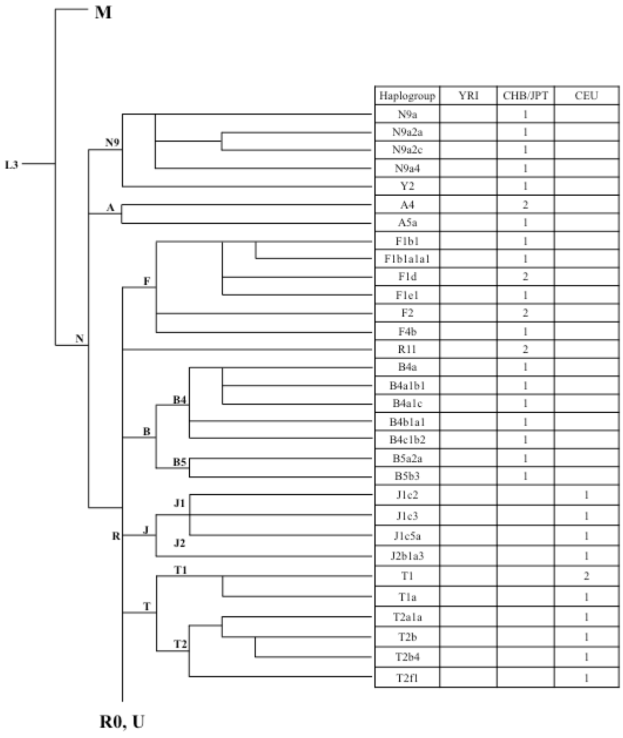
Supplementary Figure 4. Population origin of known and novel deletions in the SV discovery set. Top: Previously known deletions in the trio and low-coverage projects. Bottom: Novel deletions discovered in the trio and low-coverage projects.



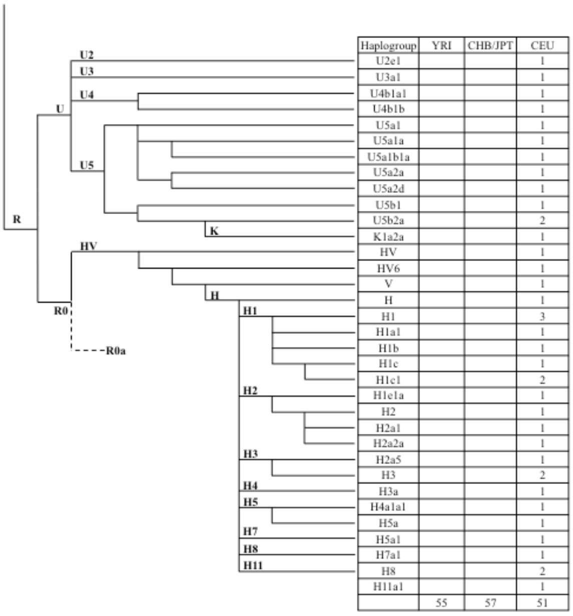
Haplogroup	YRI	CHB/JPT	CEU
L0a1a2	3		
L1b1a	3		
L1b1a3	4		
L2a1	3		
L2a1a	2		
L2a1a1	3		
L2a1c	2		
L2a1e1	1		
L2a1f	2		
L2b1	1		
L2b1a3	2		
L2b3	1		
L2c2	1		
L2e	2		
L3	3		
L3b	1		
L3b1a	3		
L3b1a1	1		
L3b2	1		
L3e	1		
L3e1a	1		
L3e1b	1		
L3e2a	1		
L3e2a1b1	1		
L3e2b	4		
L3e2b1a	1		
L3e2b2	1		
L3e2b3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

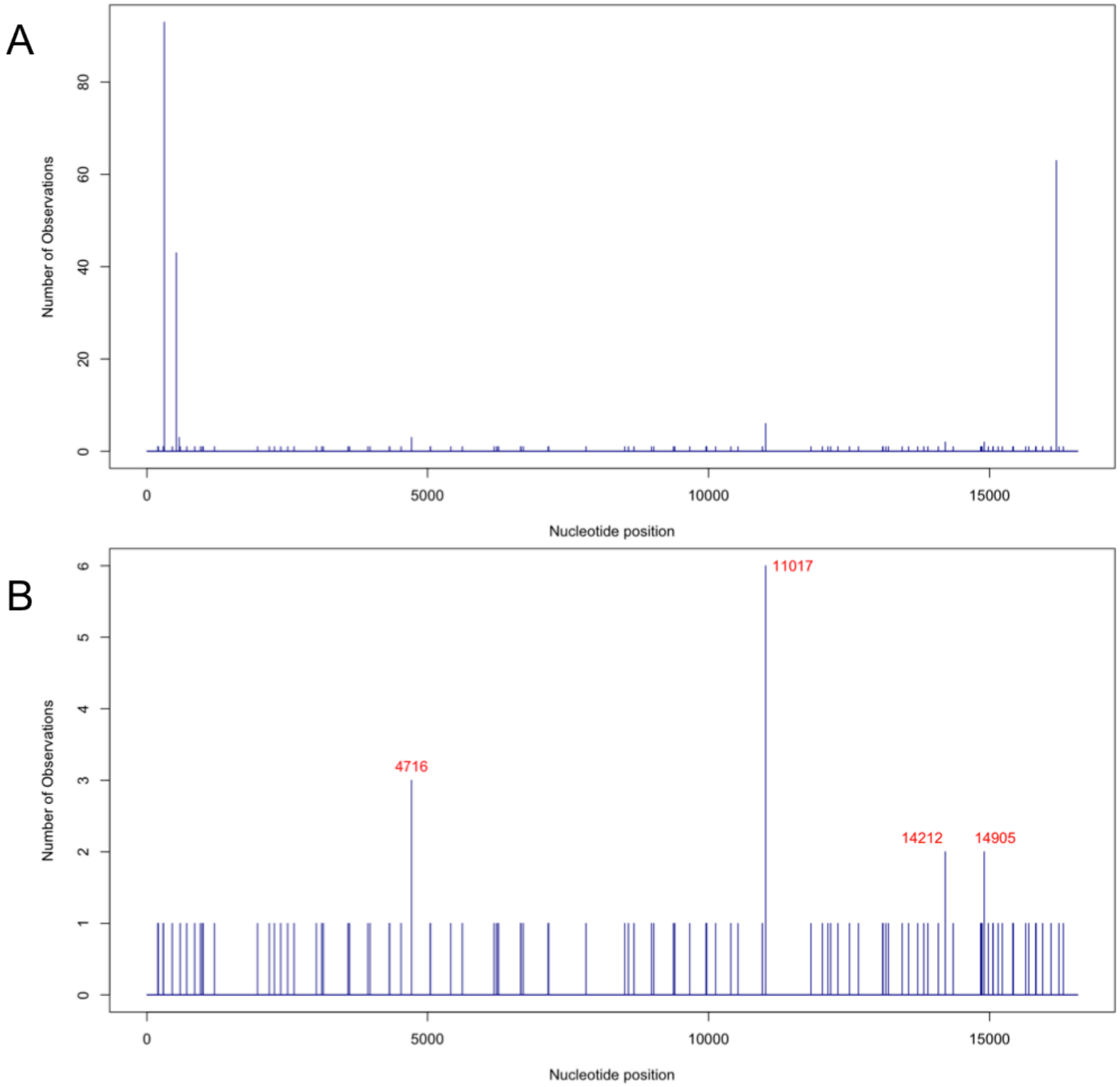


Haplogroup	YRI	CHB/JPT	CEU
N9a		1	
N9a2a		1	
N9a2c		1	
N9a4		1	
Y2		1	
A4		2	
A5a		1	
F1b1		1	
F1b1a1a1		1	
F1d		2	
F1e1		1	
F2		2	
F4b		1	
R11		2	
B4a		1	
B4a1b1		1	
B4a1c		1	
B4b1a1		1	
B4c1b2		1	
B5a2a		1	
B5b3		1	
J1c2			1
J1c3			1
J1c5a			1
J2b1a3			1
T1			2
T1a			1
T2a1a			1
T2b			1
T2b4			1
T2f1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

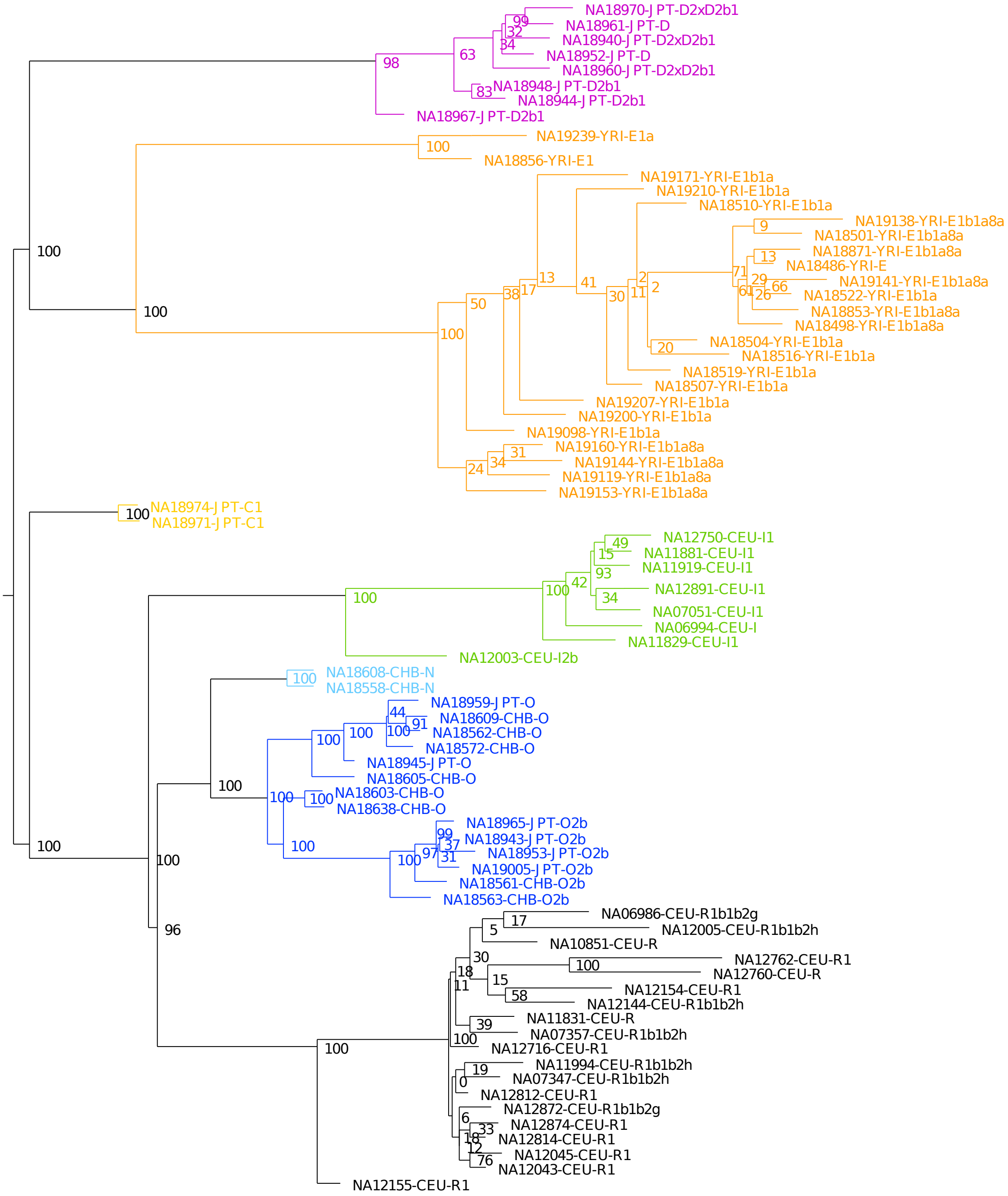


Haplogroup	YRI	CHB/JPT	CEU
U2e1			1
U3a1			1
U4b1a1			1
U4b1b			1
U5a1			1
U5a1a			1
U5a1b1a			1
U5a2a			1
U5a2d			1
U5b1			1
U5b2a			2
K1a2a			1
HV			1
HV6			1
V			1
H			1
H1			3
H1a1			1
H1b			1
H1c			1
H1c1			2
H1e1a			1
H2			1
H2a1			1
H2a2a			1
H2a5			1
H3			2
H3a			1
H4a1a1			1
H5a			1
H5a1			1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

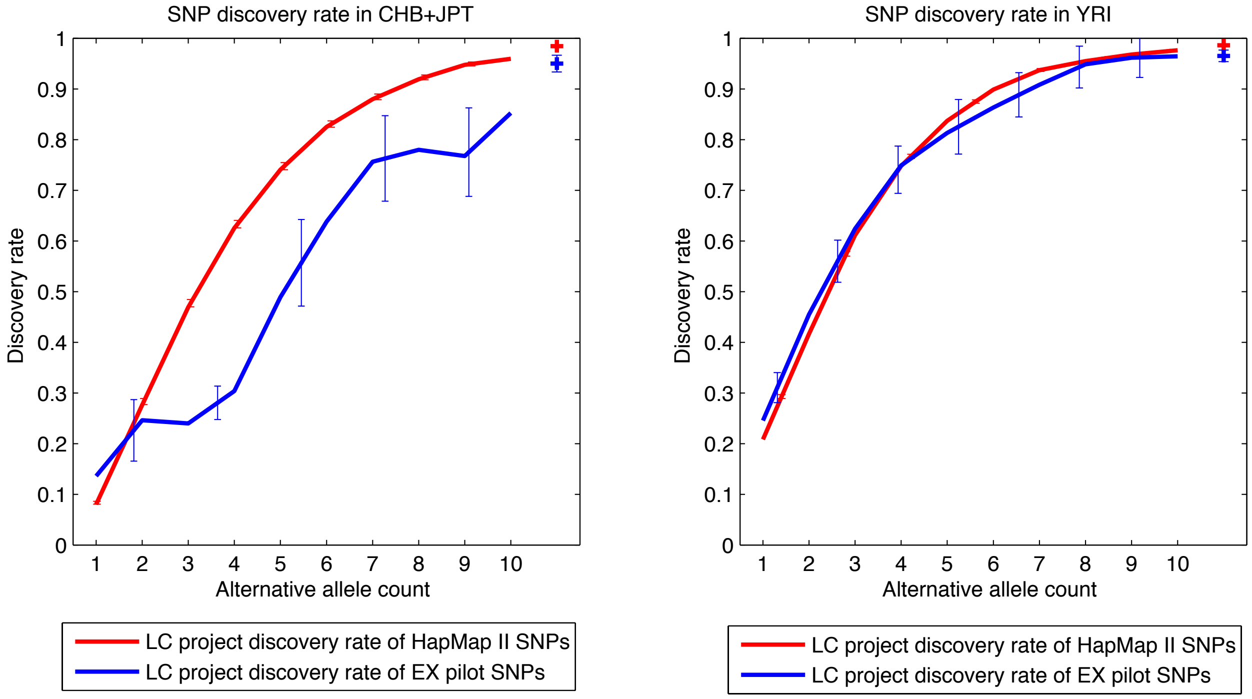
Supplementary Figure 5. Mitochondrial DNA (mtDNA) haplogroup distribution in the CEU, CHB+JPT and CEU samples in the low-coverage project. Each population sample was found to contain only previously described continent-specific haplogroups, for example haplogroup H for CEU, haplogroups D or B for the East Asian samples, and haplogroup L for YRI.



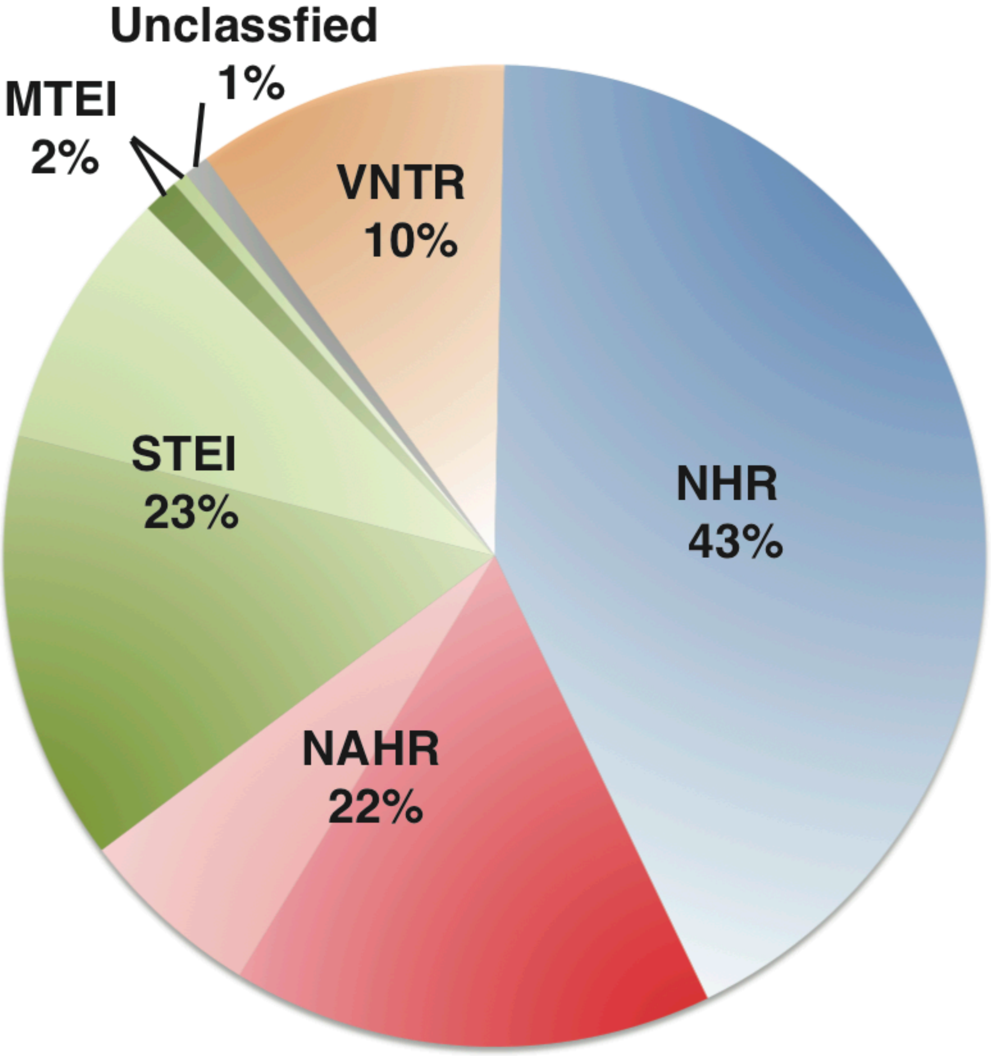
Supplementary Figure 6. Heteroplasmy variation along the mtDNA molecule (**A**) Distribution of length heteroplasmy. (**B**) Distribution of point heteroplasmy.



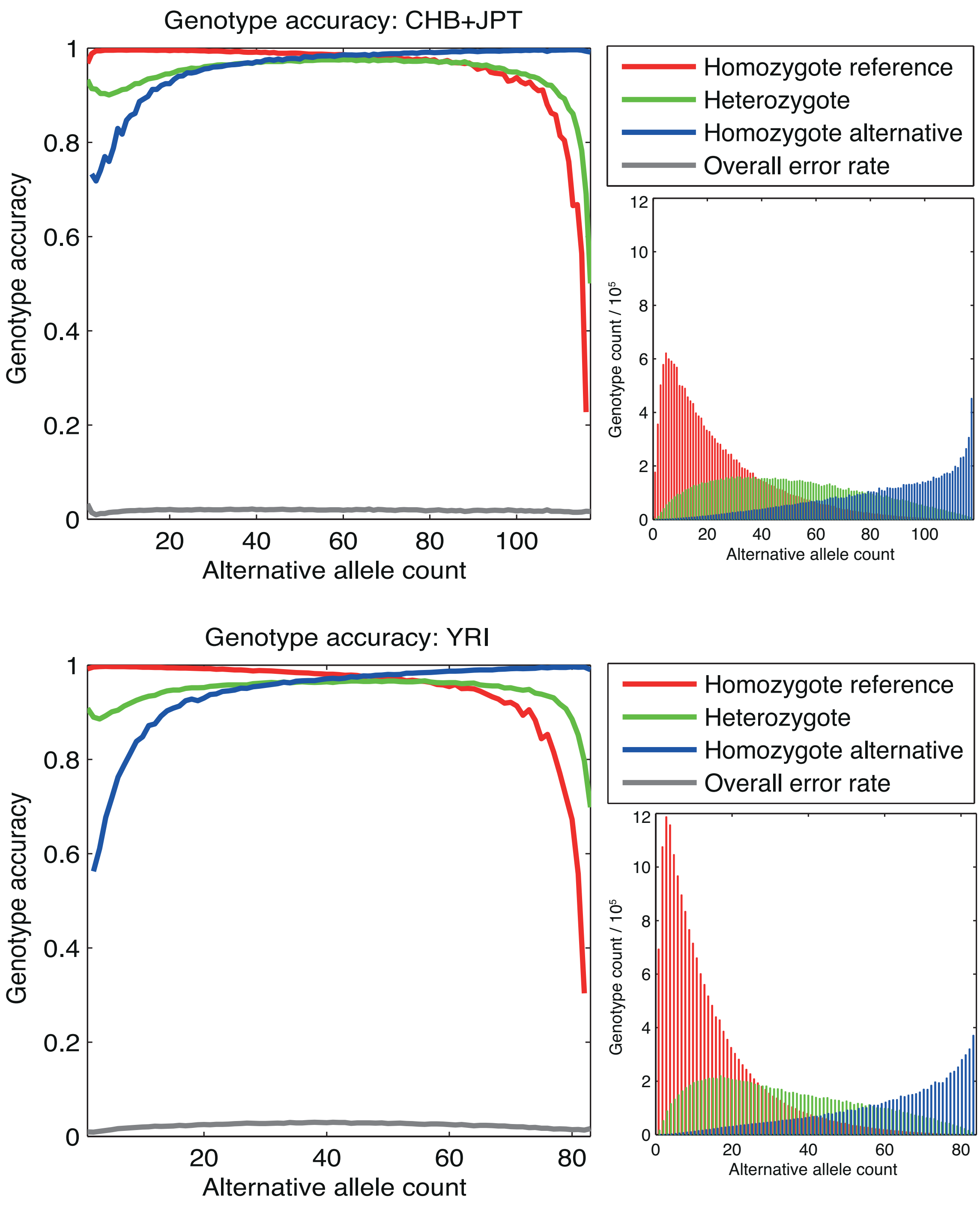
Supplementary Figure 7. The Y chromosome haplogroup tree, inferred by maximum likelihood from the 2870 variable sites identified. The leaf labels contain the population and the haplogroup assignment of each sample (also shown by colours) based on HapMap genotype data. For most haplogroups, the newly discovered SNPs gave additional resolution to the phylogenetic structure.



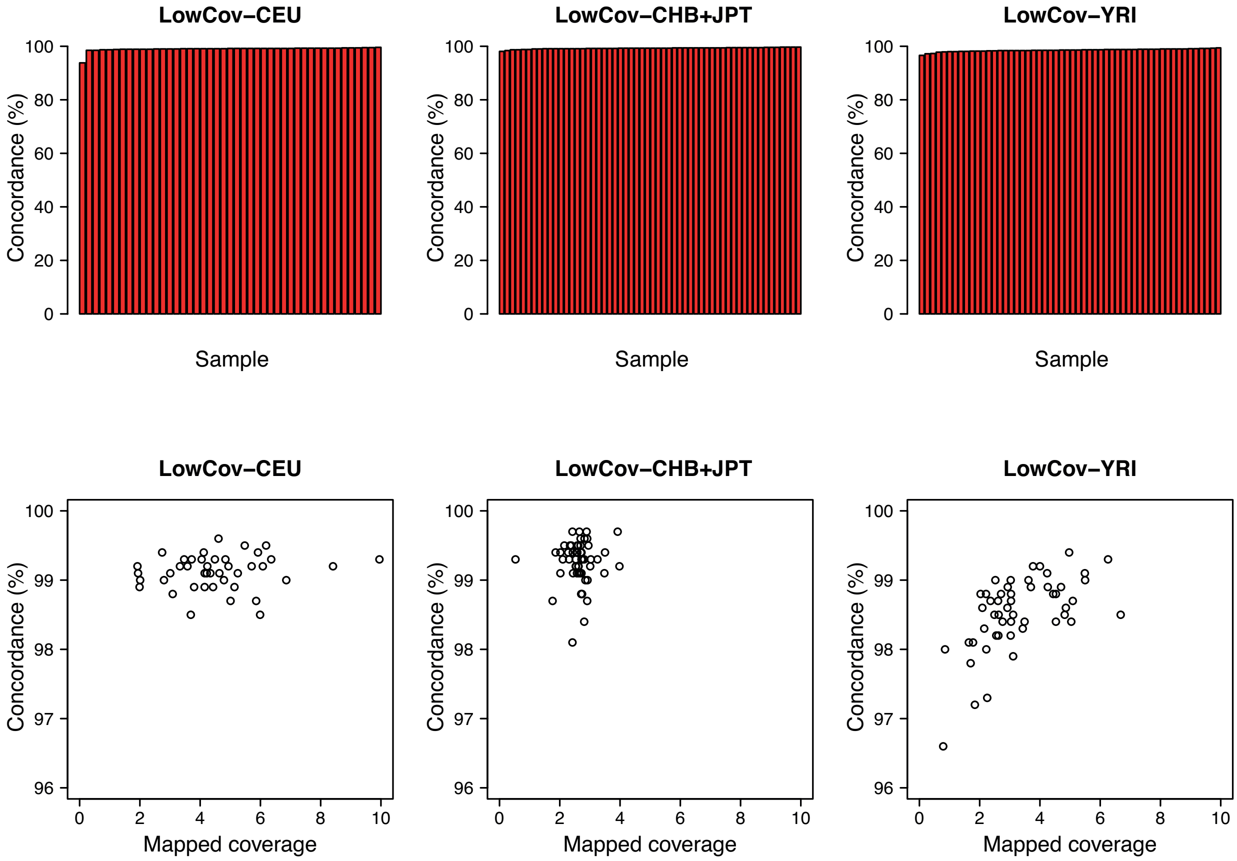
Supplementary Figure 8. Estimated power to detect SNPs in the low-coverage project for the CHB+JPT and YRI analysis panels as a function of the expected number of non-reference alleles in the sample. Crosses represent the average discovery fraction for all variants having more than 10 copies in the sample. The red lines show the proportion of HapMap II sites (excluding sites also in HapMap 3) found to be polymorphic in the low-coverage project as a function of HapMap alternative allele count. The blue lines show the proportion of exon project sites found to be polymorphic in the low-coverage project as a function of the exon project alternative allele count. For both comparisons, only samples that overlap are included. Error bars show 95% confidence intervals. Note that, as in Figure 2a, we plot power against expected allele count in the sequenced samples, e.g., a variant present in, say, 2 copies in an overlap of 30 samples is expected to be present 4 times in 60 samples.



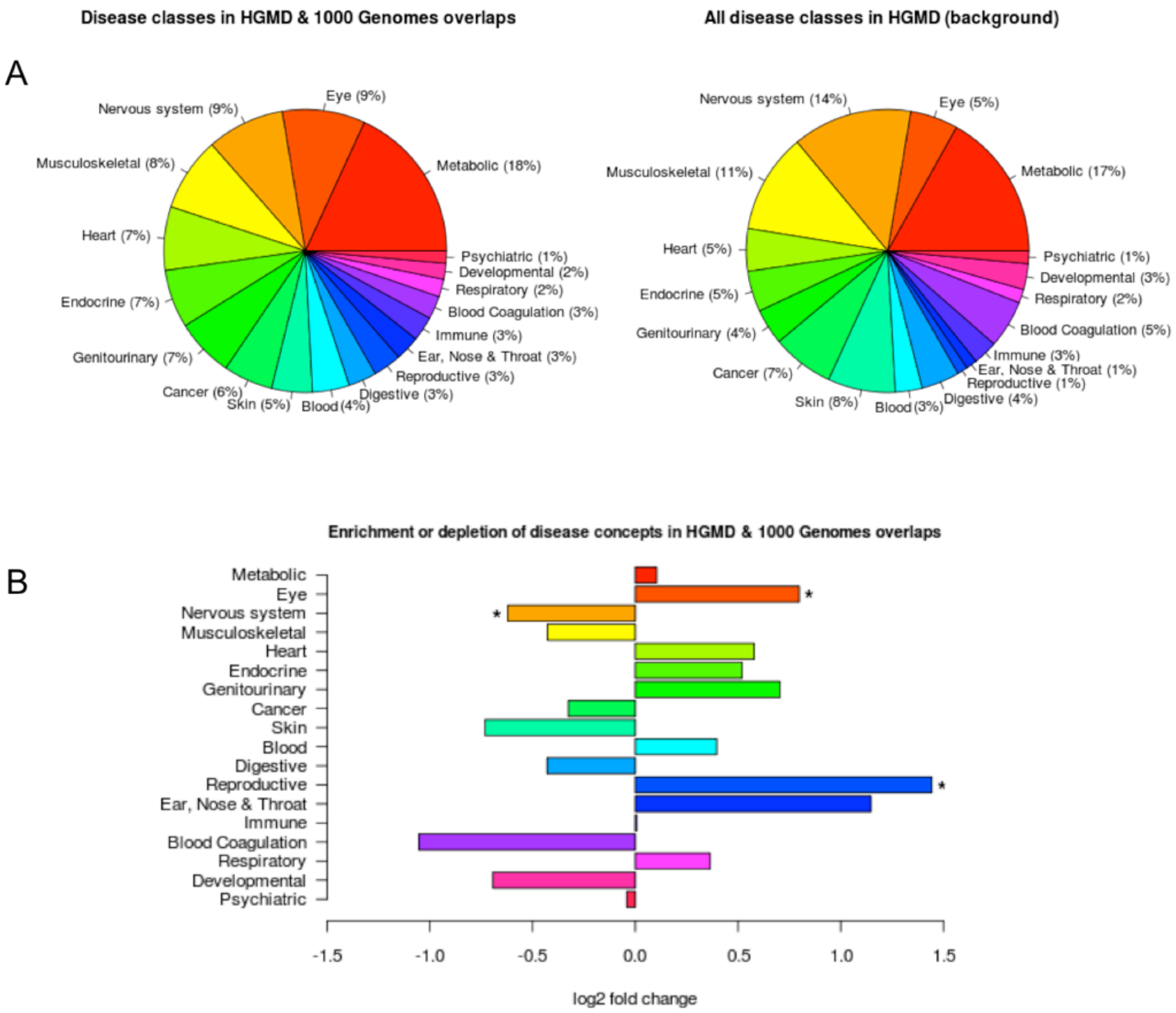
Supplementary Figure 9. Formation mechanisms of single-nucleotide resolution SVs inferred by BreakSeq (Lam, Mu et al. 2010) (see Supplementary Information Section 8.8 for details). NAHR: non-allelic homologous recombination; VNTR: variable number of tandem repeats; NHR: non-homologous end-joining (NHEJ) 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.



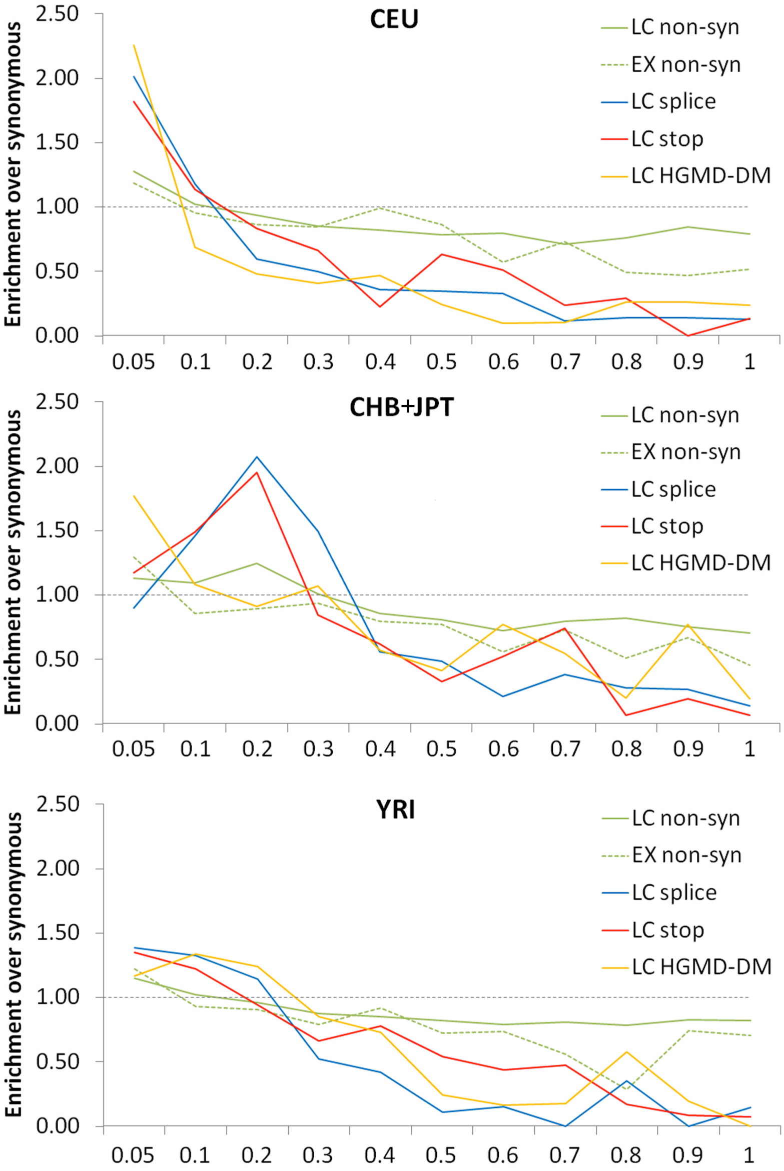
Supplementary Figure 10. Low-coverage project genotype accuracy at HapMap II sites, not found in HapMap 3, as a function of alternative allele count for the CHB+JPT (top) and YRI (bottom) analysis panels. Genotype accuracy is shown separately for homozygote reference calls (red), heterozygote calls (green), and homozygote alternative calls (blue). Also shown is the overall discordance 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.



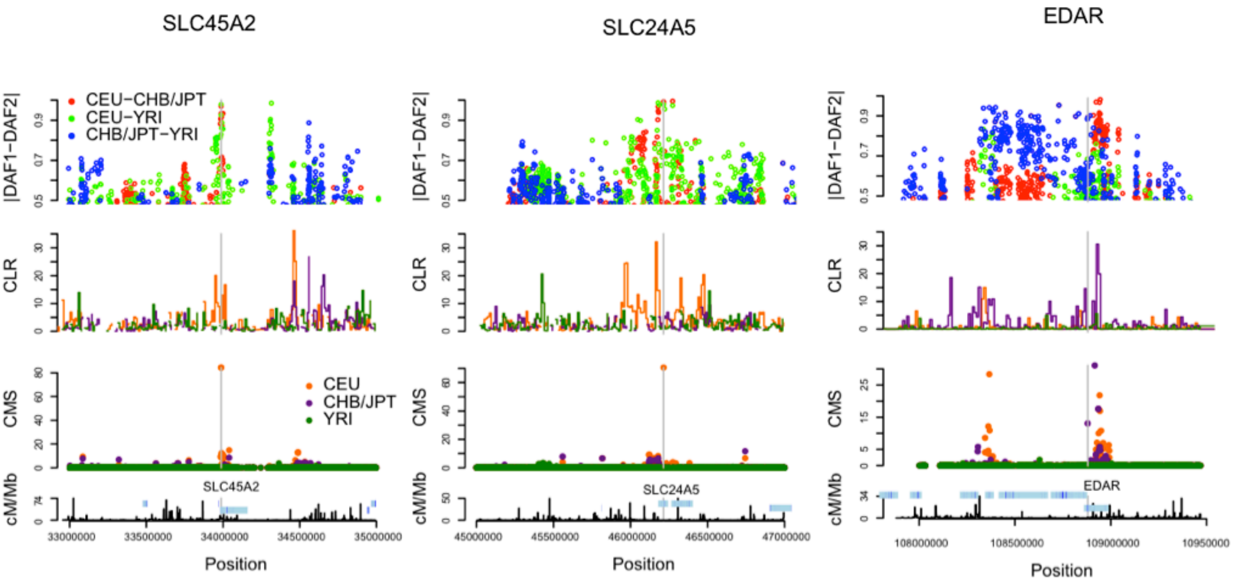
Supplementary Figure 11. Deletion genotype concordance (Top) Bar plots show the concordance of deletion genotype calls with the genotypes of Conrad *et al.* for each sample in each low-coverage analysis panel. (Bottom) Deletion genotype concordance plotted versus the mapped coverage for each low-coverage sample by analysis panel. Note that genotype concordance is consistently over 95% and typically c. 99%.



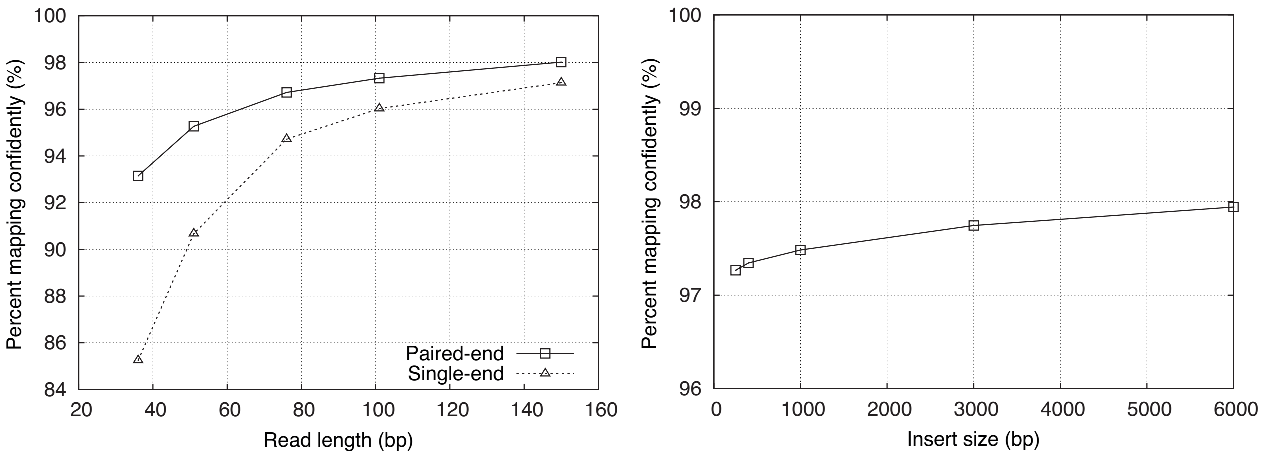
Supplementary Figure 12. (**A**) Disease class proportions in the HGMD - 1000 Genomes overlap subset (left) and HGMD background (right), labeled with class and proportion. (B) The ratio of observed to expected HGMD-DM variants found as a function of disease class, where the expected number is based on 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).



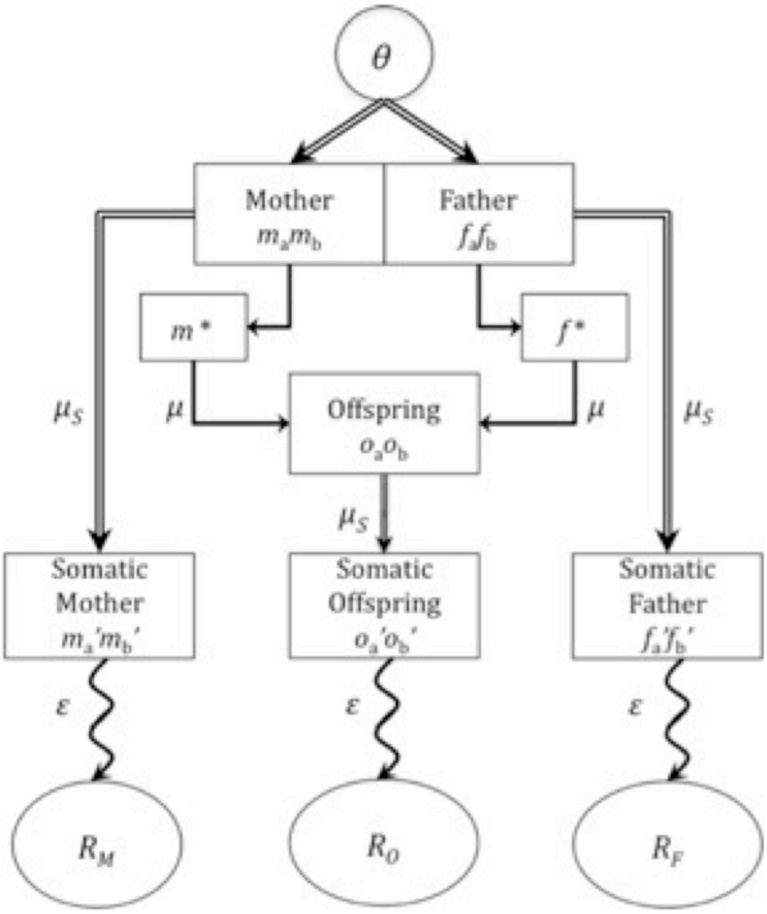
Supplementary Figure 13. Derived allele frequency spectra for different functional classes of variants, relative to putatively neutral (synonymous) coding variation. Lines indicate the relative proportion of each functional variant class in the specified frequency bin relative to the corresponding proportion for synonymous variants in the same project (i.e., low coverage or exon). 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 CHB+JPT is likely due to a higher rate of sequencing artifacts in this analysis panel, which disproportionately affect low-frequency putatively functional variation. Weaker enrichment of low-frequency putatively 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 variants 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"; LC: low-coverage project; EX: exon project.



Supplementary Figure 14. Localisation of the targets of selective sweeps around three genes previously shown to have strong signals of local adaptation and containing non-synonymous variants as candidates for the target of selection; the pigmentation genes *SLC45A2* (Sabeti, Varilly et al. 2007) (Phe374Leu at rs16891982) and *SLC24A5* (Lamason, Mohideen et al. 2005; The International HapMap Consortium 2005) (ALA111THR at rs1426654) and *EDAR* (Sabeti, Varilly et al. 2007) (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 localize signals of adaptation, the location of genes and exons (light and dark blue bars respectively) and the fine--scale recombination rate (from HapMap II). For both *SLC45A2* and *SLC24A5* the CMS statistic localizes to the non-synonymous 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 40 kb upstream of the coding variant, within the first, untranslated exon and introns 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 50 kb 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 accessibility as a function of read length with an average insert size of 400 bp (left), and as a function of insert size with an average read length of 100 bp (right). See Section 15.1 for details.



Supplementary Figure 16. Graphical representation of the statistical model used by the U de Montreal group for detecting *de novo* mutations in the trio project data. Sequencing reads covering the site of interest in the mother, father and offspring (RM,RF,RO) 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 (ma,mb) and paternal (fa,fb) 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.