An integrated map of genetic variation from 1,092 human genomes: Supplementary Material

The 1000 Genomes Project Consortium

1 Introduction ... 3
2 Materials ... 4
2.1 Criteria for choosing populations included in the project ... 4
2.2 Sample collection and distribution ... 6
2.3 Lymphoblastoid cell line establishment ... 7
3 Data generation and processing ... 8
3.1 Reuse of data from Pilot phase ... 8
3.2 Low coverage sequencing ... 8
3.2.1 Broad Institute ... 8
3.2.2 Baylor College of Medicine – Human Genome Sequencing Center ... 9
3.2.3 BGI ... 11
3.2.4 Max-Planck Institute for Molecular Genetics ... 12
3.2.5 Washington University ... 13
3.2.6 Sanger Institute ... 13
3.2.7 Illumina ... 13
3.3 Whole exome sequencing ... 14
3.3.1 Description of Exome consensus ... 14
3.3.2 Baylor College of Medicine – Human Genome Sequencing Center ... 14
3.3.3 Broad Institute ... 16
3.3.4 Washington University ... 16
3.4 OMNI genotyping ... 16
3.5 Lane level identity checks ... 17
3.6 Post-alignment identity and contamination checks ... 17
3.7 Analysis of cryptic relatedness and other sample identity checks ... 18
4 Data processing ... 18
4.1 Low coverage Illumina and 454 processing ... 18
4.2 Low coverage SOLiD read mapping ... 19
4.3 Exome Illumina read mapping ... 20
4.4 Exome SOLiD read mapping ... 20
4.5 MOSAIK low coverage and exome read mappings ... 20
5 Variant calling ... 21
5.1 Low coverage and Exome SNP calling: Broad ... 21
5.2 Low coverage SNP calling: Baylor College of Medicine HGSC ... 22
5.3 Low coverage SNP calling: University of Michigan ... 23
5.4 Low coverage SNP and indel calling: Sanger ... 24
5.5 Low coverage SNP calling: NCBI ... 24
5.6 Exome SNP calling: Baylor College of Medicine HGSC ... 24
5.7 Exome SNP calling: University of Michigan ... 25

5.8 Exome SNP calling: Weill Cornell Medical College 26
5.9 Variant calling from MOSAIK alignments 27
5.10 Creation of low coverage SNP consensus 28
5.11 Generation of consensus Exome SNP call set 29
5.12 Low coverage and exome indel calling: Broad 30
5.13 Low coverage indel calling: Sanger 31
5.14 Low coverage indel calling: Dindel2 31
5.15 Low coverage indel calling: Oxford 32
5.16 Creation of low coverage indel consensus 33
5.17 Structural variation: Deletions 34
5.17.1 BreakDancer (run at WTSI) 34
5.17.2 CNVnator 35
5.17.3 Delly 35
5.17.4 Genome STRiP 36
5.17.5 Pindel 36
5.17.6 Merging candidate deletion calls to create sensitive set 36
5.17.7 Assembly of deletion breakpoints 37
5.17.8 Performing local assembly with TIGRA-SV 37
5.17.9 Deriving breakpoints from CROSSMATCH alignments 38
5.17.10 Deriving breakpoints from AGE alignments 38
5.17.11 Classification of SV formation mechanism by BreakSeq 39
5.17.12 Creation of the specific SV discovery set and FDR estimation 39
5.17.13 Structural variation genotyping 40
5.17.14 Structural variation genotyping on chromosome Y 41
5.18 Integration of SNPs, short indels, and SVs into a single call set 42
5.19 Post-hoc short indel filtering 42
6 Variant calling on chromosome Y 43
7 Variant calling for mtDNA 45
8 Variant annotation 45
8.1 Functional annotation 45
8.2 Annotation of ancestral allele 46
9 Validation and data quality 46
9.1 Low Coverage SNP validations 46
9.1.1 Sequenom and Pacific Biosciences validation 46
9.1.2 Roche 454 validations 47
9.1.3 Consolidation of validation genotypes 48
9.1.4 Results 49
9.2 Exome validation 49
9.2.1 Exome SNP validation 49
9.2.2 Exome consensus SNPs stratified by allele frequency 49
9.2.3 Novel exome consensus SNPs 49
9.2.4 Center-unique exome SNPs 50
9.3 Loss of Function (LoF) SNP validation 50
9.3.1 LoF SNP selection 50
9.3.2 PCR-Roche 454 validations 50
9.3.3 Results 51
9.4 Short indel Validation 51
9.4.1 Sequenom and PacBio validation 52
9.4.2 454 validation 52

9.4.3 Results of PCR-based validation ... 52
9.4.4 Axiom Exome genotyping array ... 53
9.4.5 Short indel filtering ... 53
9.5 SV validation ... 54
9.5.1 SV validation using Omni 2.5 SNP genotyping arrays ... 54
9.5.2 SV validation using PCR ... 54
9.5.3 SV validation using custom CGH microarrays ... 56
10 Analysis ... 57
10.1 Quantifying the Phase 1 dataset ... 57
10.2 Assessment of power of variant discovery and genotype accuracy ... 58
10.3 Variant discovery by low coverage and exome sequencing ... 59
10.4 Assessment of the accessible genome in Phase 1 ... 59
10.5 Haplotype estimation from OMNI data ... 60
10.6 Imputation using the Phase 1 data ... 61
10.6.1 SNP and indel evaluation with Complete Genomics data ... 61
10.6.2 SV evaluation with Conrad *et al.* data ... 62
10.6.3 Comparison of Phase 1 haplotypes with benchmark haplotypes ... 63
10.7 Analysis of private and cosmopolitan variants by frequency ... 64
10.8 Density of variants as a function of derived allele frequency ... 65
10.9 Analysis of highly differentiated sites ... 65
10.10 Rare allele sharing within and between populations ... 66
10.11 Shared haplotype length as a function of allele frequency ... 67
10.12 Local ancestry inference ... 67
10.13 Estimation of F_{ST} ... 69
10.13.1 Weir and Cockerham, HapMap estimators ... 69
10.13.2 Hudson ratio of averages ... 69
10.14 Quantifying potentially functional variants in the Phase 1 dataset ... 70
10.14.1 Coding variant classes ... 70
10.14.2 Identification and filtering of loss-of-function variants ... 70
10.14.3 HGMD-DM and COSMIC SNPs ... 71
10.14.4 Non-coding variant classes ... 71
10.14.5 Other conserved variants ... 72
10.15 Rare variant proportion for variants with functional consequences .. 72
10.16 Estimation of excess nonsynonymous variants in KEGG pathways ... 72
10.17 Nucleotide diversity around CTCF binding motifs ... 73
10.18 Population differentiation of functional SNPs ... 73
10.19 Variants in linkage disequilibrium with focal GWAS SNPs ... 74
10.20 Comparison of 1000 Genomes Phase 1 to UK10K study ... 74
10.21 Comparison of 1000 Genomes Phase 1 to SardiNIA study ... 74
11 Accessing 1000 Genomes data ... 75
12 References ... 75

1 Introduction

In this Supplementary Text we give further technical information regarding the 1000 Genomes Phase 1 data collection, processing, validation, and analysis. The aim is to record in more detail than is possible in the main text how the call sets were generated and analyzed.

The 1000 Genomes Phase 1 release is the result of a large number of people working in collaboration. Where possible, we have identified individuals associated with each section of the supplement in order to provide the reader a means of identifying individuals contributing to each area of the project.

2 Materials

2.1 Criteria for choosing populations included in the project

Authors: Aravinda Chakravarti, Bartha Knoppers, Lisa Brooks, and Jean McEwen

The choices of the specific populations to include in the project were informed by recommendations made by the project's Samples and ELSI Group, and were based on a mix of scientific, ethical, and practical considerations. The most important scientific rationale was to expand the sampling of humans so that within each continent we had multiple yet distinct populations. The underlying objectives were to obtain genetic variation data that would be broadly representative for the vast majority of individuals within a continent (although we did not attempt to cover most of the diversity in Africa, since there is too much for the sample sizes that could be done for this project). We are well aware that this is a great challenge since the continents differ greatly in the amount and patterns of their internal genetic diversity

The criteria for inclusion were:

Broad consent: The participants had to have provided consent for broad use of the samples and data, and for broad data release in databases available on the Internet.

No names or other traditional individual identifiers: To protect the privacy of the research participants, no names or other traditional individual identifiers were collected. For a few of the populations, where recruitment was carried out in conjunction with an ongoing genetic study, names that had previously been collected were retained by the original investigators in strict confidence but have not been shared with this project.

No phenotype data: To protect the privacy of the research participants, the project preferred that no phenotype or clinical data be collected. For a few of the populations, where recruitment was carried out in conjunction with an ongoing genetic study, such data had previously been collected and are available only to the sample collectors and their authorized collaborators, and not to the project.

Cell lines, samples available to many researchers for many future uses: Research participants must have consented to have cell lines made from their samples, to have the cell lines stored in a public repository, and to have the cell lines and DNA from the cell lines distributed to a broad range of researchers for use in a wide range of genetic and genomic studies, including studies of molecular phenotypes such as gene expression and response to drugs.

Trios: The project wanted to include samples from mother-father-adult child trios wherever possible since even a few trios for a population helps in assessing the quality of the data, confirming rare variants, and inferring haplotypes. However, at many of the collection sites it was impossible to obtain samples from all three members of trios within the project timeline. For these populations, the samples include a mix of trios, parent-child duos, and unrelated individuals.

Genomic data already available: The project wanted to include some samples on which considerable data were already available, such as the HapMap Project samples. This allows comparisons with other sets of genotype, sequence, and array data on these samples, for validation of the project data and efficacy of quality control procedures.

GWAS studies: Where possible, the project sought to include samples from populations in which GWAS studies were already being conducted. Sampling in most cases was conducted in collaboration with a research center or hospital, where a relationship of trust with the community had already been established. This minimized the risks of misunderstanding of the research and increased the likelihood that the data will eventually provide some scientific benefit for the studied populations through the GWAS studies.

Large populations, not anthropological sampling: To minimize the risk of stigmatization or breach of privacy, particularly in populations that may be vulnerable because of small size, the populations chosen for sampling had to be reasonably large. The precise way that each sampled population was defined depended on the locality. In some places a single population, defined by ethnicity, was collected, while at other places the samples came from the geographic region or country with no emphasis on being from a particular ethnic or ancestral group. The goal was not to define populations in an anthropological sense, but to collect samples that addressed the project's biomedical goals while recognizing the complexities of local populations and how they define themselves.

As a privacy safeguard, more samples were collected from each population than were actually studied for the project. In this way it is unknown whether the sample from any particular sampled person was actually used in the project.

Eight of the HapMap I/II and HapMap 3 population samples met the above inclusion criteria and were included in the project:

- People with African Ancestry in the Southwest United States (ASW)
- Han Chinese in Beijing, China (CHB)
- Japanese in Tokyo, Japan (JPT)

- Luhya in Webuye, Kenya (LWK)
- People with Mexican Ancestry in Los Angeles, California (MXL)
- Utah residents with ancestry from Northern and Western European, US (CEU)
- Toscani in Italia (TSI)
- Yoruba in Ibadan, Nigeria (YRI)

One other set of existing samples, collected from the Finnish in Finland (FIN) by the late Leena Peltonen of the Wellcome Trust Sanger Institute, Hinxton, UK and the University of Helsinki, Helsinki, Finland, met the above inclusion criteria and were also included.

Samples from six additional populations were collected as part of the sampling plan. These were:

- British from England and Scotland, UK (GBR)
- Colombians in Medellin, Colombia (CLM)
- Han Chinese South, China (CHS)
- Iberian Populations in Spain (IBS)
- Puerto Ricans in Puerto Rico (PUR)

More detail about each population can be found on the website of the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute[1]. This includes information about how each population was defined for the project, the rationale for the shorthand label chosen for each population, and other important background information relating to each population and the specific sampling strategy used for that population.

The way that a population is named in studies of genetic variation such as the 1000 Genomes Project has important ramifications scientifically, culturally, and ethically. It is thus important to use care in labeling the populations when publishing or presenting the findings of studies that use project data. Guidelines on how to refer to the populations can be found online[2].

Samples from additional populations will be included in later phases of the project.

2.2 Sample collection and distribution

Authors: Aravinda Chakravarti, Bartha Knoppers, Lisa Brooks, and Jean McEwen

Each research group that collected samples from a population first provided a detailed written sampling plan, for the Samples and ELSI Group to review and approve. Each sampling plan included information about the social and demographic characteristics of the population proposed for sampling and the specific community where recruitment was to occur; information about the informed consent process; and information about any particular concerns anticipated to arise in the community and how they would be addressed. The research groups obtained all required IRB and ethics approvals from their institutions and collaborating institutions, as well as government approvals where required. Few major issues arose during the course of sampling that the Samples and ELSI Group was asked to address, but the group remained available throughout the course of the sampling to consult on any matters that did arise.

The samples are stored at the NHGRI Sample Repository for Human Genetic Research at the non-profit Coriell Institute for Medical Research in Camden, New Jersey (Coriell). When Coriell receives orders from researchers for samples, it screens the statements of intended research use. Coriell provides regular reports to each community where sampling occurred on the use of the samples from that community, and has funds available to support educational or outreach activities related to biomedical research in the communities. Coriell provides each research group that collected samples in a community with a free set of DNA or cell lines from those samples. Coriell charges researchers in resource-limited countries a much reduced rate for sample DNA.

2.3 Lymphoblastoid cell line establishment

Authors: Neda Gharani and Lorraine H. Toji

Lymphoblastoid cell cultures were established at the Coriell Cell Repositories from fresh bloods after separating the mononuclear cells on a Ficoll gradient and incubating with Epstein Barr virus and phytohemaglutinin in RPMI 1650 with 15% v/v fetal bovine serum[3].

When a transformed cell culture was obtained, sufficient cells were grown to cryopreserve 40 to 60 ampoules at 5 million cells per ampoule; 8 to 10 amps of these are reserved for future expansion to replenish cell culture and DNA distribution stocks. The remainder are available for distribution as cell cultures to investigators around the world. As part of the cell culture quality control, cultures are tested for sterility and confirmed to be free of mycoplasma, bacteria, and fungi[4]. Frozen LCLs are also checked for viability by checking growth of a recovered ampoule of frozen cells. In addition, LCLs are screened for presence of HIV proviral sequences. Quality control[5] to detect possible misidentification of samples is carried out by comparing each cell culture expansion and each lot of DNA to the original submission using a set of six highly polymorphic microsatellite markers (supplemented by the ABI Amplifier panel to resolve ambiguities) and an amelogenin gender assay; these data are also used to confirm family relationships of trios.

From some populations (GBR, FIN and IBS) one or two ampoules of frozen lymphoblastoid cell cultures, established elsewhere, were submitted to the Repository. Frozen LCLs were cultured, expanded to the required cell numbers to create distribution and reserve cell culture stocks that were subjected to the same cell culture quality control tests as above. Because no original blood was available for these samples, the identity quality control relied on consistency of family relationships (if trios were collected) and gender information provided by the submitting group. A portion of each frozen culture stock is reserved for replenishment of cell culture stocks and DNA.

Therefore, for as long as possible, replenishment of distribution stocks of cell cultures and DNA goes back to the original frozen cell culture stock. If the original cell culture stock is ultimately depleted, the reserved amps of an expansion of that original stock will become the new reserve stock and will be approximately 5 to 7 population doublings beyond the original culture stock.

3 Data generation and processing

3.1 Reuse of data from Pilot phase

Authors: Laura Clarke, Xiangqun Zheng-Bradley, and Richard E. Smith

The pilot phase of the 1000 Genomes Project[6] consisted of three separate projects: low coverage whole genome sequencing of unrelated individuals; deep coverage whole genome sequencing of two family trios; and exon targeted sequencing for a subset of approximately 1000 genes. Phase 1 of the 1000 Genomes Project continued with the low coverage sequencing strategy of the Pilot phase of the project, and also conducted whole exome sequencing on all samples. Trio and exon targeted sequence data from the Pilot was not included in Phase 1. However, the low coverage Pilot data has been included as part of Phase 1.

For Phase 1, we performed Quality Control (QC) and filtering of the input FASTA files available in the Short Read Archive (SRA)[7]. This QC and filtering consists of a series of syntax and sequence checks to ensure the data meets minimum formatting and quality criteria.

The Pilot phase contained early data from the Illumina, 454 and SOLiD platforms. Much of the early SOLID data consisted of 25 bp reads that showed substantial reference bias. The Project decided to remove this data, rather than attempt adjustment of analysis methods to correct for this bias.

3.2 Low coverage sequencing

3.2.1 Broad Institute

Author: Namrata Gupta

Libraries were constructed then sequenced on either an Illumina HiSeq 2000 or Illumina GAIIX with the use of 101 bp paired-end reads. Output from Illumina software was processed by the Picard[8] data-processing pipeline to yield BAM files containing well-calibrated, aligned reads. All sample information tracking was performed by automated LIMS messaging.

For a subset of samples, starting with 3µg of genomic DNA, library construction was performed as described by Fisher *et al*[9]. Another subset of samples, however, was prepared using the Fisher *et al*. protocol with some slight modifications. Initial genomic DNA input into shearing was reduced from 3µg to 100ng in 50µL of solution. In addition, for adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters containing unique 8 base index sequences embedded within the adapter.

For a subset of samples, size selection was performed using gel electrophoresis, with a target insert size of either 340bp or 370bp +/- 10%. Multiple gel cuts were used for libraries that required high sequencing coverage. For another subset of samples, size selection was performed using Sage's Pippin Prep.

Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH using Perkin-Elmer's MultiProbe liquid handling platform. The subset of the samples prepared using forked, indexed adapters was quantified using qPCR, normalized to 2nM using Perkin-Elmer's Mini-Janus liquid handling platform, and pooled by equal volume using the Agilent Bravo. Pools were then denatured using 0.1 N NaOH. Denatured samples were diluted into strip tubes using the Perkin-Elmer MultiProbe.

Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using either Genome Analyzer v3, Genome Analyzer v4, HiSeq 2000 v2, or HiSeq v3 cluster chemistry and flowcells. For a subset of samples, after cluster amplification, SYBR Green dye was added to all flowcell lanes, and a portion of each lane visualized using a light microscope, in order to confirm target cluster density. Flowcells were sequenced either on Genome Analyzer IIX using v3 or v4 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.7.48; or on HiSeq 2000 using HiSeq 2000 v2 or v3 Sequencing-by-Synthesis Kits, then analyzed using RTA v1.10.15 or RTA v.1.12.4.2. For whole genome sequencing, 101 bp paired-end reads were used. For pooled libraries prepared using forked, indexed adapters, Illumina's Multiplexing Sequencing Primer Kit was used, and a third, 8 base sequencing read was performed to read molecular indices.

3.2.2 Baylor College of Medicine – Human Genome Sequencing Center

Author: Donna Muzny

SOLiD Mate Pair Libraries and Sequencing Methods

SOLiD 2 x 50 bp library construction preparation was performed using standard reagents (SOLiD mate pair Library Oligo kit [LMP], 4400468) and protocol as specified by Applied Biosystems (SOLiD System 3.0 Mate-Paired library Preparation Guide). For this protocol, 20 ug of genomic DNA was fragmented by HydroShear (Digilab Genomic Solutions Inc) to an average size of 1.5 kb fragments. The fragmented DNA was repaired using the End-It DNA End Repair Kit (Epicentre) followed by ligation of the LMP CAP adaptor. Size selection of the ligation product was then performed using 1.0% agarose gel electrophoresis for resolution and 1-2 kb DNA fragments were excised from the gel and purified using QIAquick Gel Extraction Kit (Qiagen). The DNA circularization reaction was catalyzed using the Quick Ligation kit (New England Biolabs) with SOLiD biotinylated Internal Adaptors (approximately 10-15 pmoles, based on the quantity of the size-selected DNA). Following circularization, DNA Plasmid-Safe ATP-Dependent DNase (Epicentre) was used to eliminate un-circularized DNA, resulting in 500 ng – 1 ug of circularized DNA product after purification. The circularized DNA was then nick translated using DNA Polymerase I (New England Biolabs) at 50-100 units/reaction, depending on the quantity of the circularized product. Following nick translation and purification, the DNA fragments were sequentially digested with 10-20 units of T7 exonuclease (New England Biolabs) and S1 nuclease (Invitrogen). The digested DNA product was end-repaired using the Epicenter End-It DNA End Repair Kit and was bound to the Dyna MyOne C1 Streptavidin beads (Invitrogen) through the biotin-labeled internal adaptor. P1 and P2 adaptors were ligated as specified by the LMP protocol to the library fragments, and the ligated product was PCR amplified using SOLiD Library PCR Master Mix with minimum cycling (5-7 cycles) to maintain library complexity. The library was size selected on a 4% agarose gel to obtain 250-350 bp fragments. Final QC of the library was performed using a Picogreen assay (Invitrogen) and Agilent Bioanalyzer 2100 trace with DNA 7500 Chip.

SOLiD Sequencing Methods

The mate-pair libraries were clonally amplified onto 1 um beads using emulsion PCR with a final library concentration of 0.70 to 0.85 pM. Emulsion PCR reactions were processed using either the Life Technologies Full-Scale ePCR reaction protocol or a modified version of the Macro-Scale ePCR reaction protocol. As specified by the vendor, the full-scale reactions used the IKA Ultra-Turrax to generate the emulsions followed by amplification with standard thermal cycling methods. The 4X macro-scale emulsions were generated using a Servodyne Electronic Mixer (Cole-Parmer, EW-50008-30, EW-50008-00) at a speed of 780 rpm for 20 min. The 4X bulk reaction was then amplified in a sealable bag using a Hydrocycler (K-Biosciences, HC-16) with the following cycling conditions: denature for 10 min at 95^{o}C, followed by 40 cycles of 1 min at 95^{o}C, 2 min at 62^{o}C and 2 min at 72^{o}C with a final extension of 10 min at 72^{o}C. Beads were recovered by centrifugation with 2-butanol in 50 ml conical centrifuge tubes and then enriched and 3’ modified according to the Life Technologies Macro-Scale ePCR reaction protocol. The 3’ modified template positive beads were deposited and covalently bound to SOLiD sequencing slides and then underwent sequencing using the 2x50bp run format on either the SOLiD V3, V3+, or V4 platform. Sequencing was performed according to the manufacturer’s protocols and using the Life Technologies SOLiD MP Library Sequencing Kit (4406398), Opti MP Library Sequencing Kit (4442058), and ToP MP Library Sequencing Kit (4452685) for the V3, V3+, and V4 platforms respectively.

SOLiD Primary Data Processing and Sequencing QC

Base and quality calling for the SOLiD data was performed on-instrument using standard vendor software and settings. Upon completion of a run, read and quality data were copied into our data-center where individual sequence events were split into 10M read bundles to undergo preliminary quality control mapping using BFAST. After read group bundles were mapped, their results were merged back into a single sequence-event-level BAM, and where necessary, these BAMs were merged into a sample-level BAM using Picard, and duplicate reads were marked at the library level using SAMtools. Alignment metrics and uniqueness were evaluated to confirm that the sequencing performed as expected and to verify that each sample met a minimum of 6X sequencing coverage. In addition, sample concordance analysis was also performed by comparing SNP array genotypes to the sequencing data to confirm sample identity and evaluate contamination.

3.2.3 BGI

Author: Jun Wang

Genomic DNA for all Phase I samples of 1000 Genomes Project was obtained from Coriell Institute for Medical Research and was sequenced from May 2008 to October 2010. The sequencing was carried out on mainly two platforms: Illumina (Genome Analyzer and HiSeq 2000) and Applied Biosystems (SOLiD). For each sample, the data coverage was at least 4X.

For Illumina sequencing platform, library preparation complied with the Illumina's instruction. In short, 2-5 ug of genomic DNA was fragmented by nebulization with compressed nitrogen gas. After adding "A" base and DNA adaptors to the blunt DNA fragments, DNA products were then separated on a 2% agarose gel, excised from the gel at a position between 150 and 250 bp (450 and 550 bp from 2009), and purified (Qiagen Gel Extraction Kit). The modified DNA fragments were enriched by PCR with PCR primers 1.1 and 2.1 (Illumina). The concentration of the libraries was measured by absorbance at 260nm. The libraries were hybridized to the flow cells of Genome Analyzer/Genome Analyzer IIx/HiSeq 2000 for paired-end sequencing. The fluorescent images were converted to sequence using the Illumina base-calling pipeline (Solexa, 0.2-2.6; HiSeq 2000, 1.0-1.3). The length of obtained read contained 44 bp, 75 bp and 90 bp.

For SOLiD sequencing platform, genomic DNA was taken to construct libraries, which included two type, mate-pair libraries (~20 ug) and pair-end libraries (~3ug), and the ranges of insert size were 1.5-2 kb and 150-200 bp respectively. In accordance with the manufacturer's protocol (Applied Biosystems SOLiD Library Preparation Protocol), in brief, library preparation, emulsion PCR, slide preparation and sequencing were all performed. SOLiD sequencing libraries were amplified by a limited PCR reaction (mate-pair, < 12 cycles; pair-end, < 8 cycles) with a Hi-Fidelity PCR Supermix (Invitrogen, 12531-016). The length-fixed PCR band was excised from a 2% agarose gel, purified with the Qiaquick gel extraction kit (Qiagen, 28706) and quantified with the Agilent 2100 Bioanalyzer. Sequencing was carried out on the SOLiD system (V2.0 and V4.0). Image analysis and base-calling were carried out with the Applied Biosystems pipeline (SOLiD V2.0, Corona lite v4.2; SOLiD V4.0, BioScope v1.2/v1.3).

3.2.4 Max-Planck Institute for Molecular Genetics

*Authors: Marc Sultan, Marie-Laure Yaspo, and Ralf Sudbrak**

** Corresponding Author*

Genomic DNA sequencing of samples of the Phase 1 of the 1000 Genomes project was fulfilled between July 2009 and October 2010 on one of two sequencing platforms: Genome Analyser II (GAII, Illumina) and SOLiD versions 3-4 (Applied Biosystems).

For the Illumina platform, 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 (from March 2009 – 50 bp PE runs) were performed for each flowcell, allowing identification of 36 (or 50) nucleotides from each side of the genomic DNA insert.

For the SOLiD sequencing platform, mate-pair libraries (1.5-2kb) were prepared using the ABI protocol with several modifications. Shearing was performed on Hydroshear. To avoid chimeras additional size selection after shearing was performed. For all purification steps except for gel purification phenol/chloroform purification was used. In the circularization reaction the ratio DNA molecules : Internal adapter was changed to 1: 1.2; ligation time was increased to 1.5 hours (or more). For test amplification and large-scale amplification not Invitrogen mix, but 2x Phusion HF Master Mix (NEB, #F-531L) was used. Resulting beads with attached library molecules were loaded onto the flowcell (amount of usable beads varied from 450 to 500 million per single-frame flowcell). For each flowcell, 50 bp mate pair run was performed.

For both platforms raw data was pipelined according to corresponding manufacturer's instructions. Illumina's Genome Analyzer Sequencing Control Software (SCS) v2.4 and SOLiD Analysis Tools pipeline were used for base calling for Illumina and SOLiD, respectively. For preliminary analysis, resulting sequencing reads were aligned to the human genome (hg18, NCBI build 36.1). The identity of each sample was confirmed using HapMap genotypes.

3.2.5 Washington University

Author: Elaine Mardis

Genomic DNA was obtained from the Coriell Institute for Medical Research. Illumina libraries were constructed according to the manufacturer's (Illumina Inc, San Diego, CA), recommendations with the following exceptions: 1) 500 ng of DNA was sheared using a Covaris S220 DNA Sonicator (Covaris, INC. Woburn, MA) to a size range between 200-500 bp. 2) PCR optimization was performed to determine the optimal cycle number to prevent over-amplification, thus decreasing duplication rates. 3) Eight PCR reactions were amplified to enrich for proper adaptor ligated fragments. 4) The final size selection of the library was achieved by electrophoresis of the enriched library on a 4-10% PAGE gel, and isolating 50 to 100 bp fractions within a window size of 350-400 bp, 300-400 bp or 450-500 bp. The fractions collected include the Illumina adaptor sequences. qPCR was used to determine library concentrations. Libraries were sequenced on the Illumina GAIIx with a target of greater than 4X coverage of the human genome. Four lanes of 2 X 101 bp read pairs were generated for each sample with a minimum of 12 Gb per sample as a minimum passing criterion.

3.2.6 Sanger Institute

Authors: Thomas Keane and Jim Stalker

Genomic DNA was obtained from the Coriell Institute for Medical Research. Three different types of libraries were produced: standard, high-complexity, and noPCR. For all libraries we began with 5 µg of genomic DNA. The standard libraries were constructed according to the manufacturer's recommendations (Illumina Inc, San Diego, CA). Each sample was fragmented using a disposable nebulizer (Invitrogen) and purified using a qiaquick column (Qiagen). DNA was end-repaired and adaptors ligated to the ends of the DNA. Fragments of approximately 300-400 bp were gel-purified and PCR amplified. For details of the high-complexity and noPCR library preparation adaptations, see Quail *et al.*[10]. Flow cells were prepared, clusters generated, and processed flowcells were paired-end sequenced from each end on either an Illumina Genome Analyzer II (108 bp) or HiSeq 2000 (100 bp). For each lane, reads were aligned to hg19/NCBI37 and HapMap genotype validation was performed.

3.2.7 Illumina

Author: Sean Humphray

Genomic DNA was acquired from the Coriell Institute for Medical Research. The samples were sequenced between March and May 2010. Preparation of short-insert paired-end Illumina sequencing libraries, flow cell preparation and cluster generation was conducted as has been described previously[11]. Briefly, genomic DNA samples (4 µg) were randomly fragmented by nebulization and used to prepare paired-end sequencing libraries with average insert size of 303 bp human insert and a 6% 1xSD. Libraries were denatured using NaOH (0.1 N) and diluted in cold (4 °C) hybridisation buffer (5x SSC + 0.05 % Tween 20) prior to seeding clusters on the surface of the flow cell. Cluster amplification, linearization, blocking and hybridisation to the Read 1 sequencing primer were carried out on a Cluster Station. Following the first sequencing read, flow cells were held *in situ* and clusters were prepared for Read2 sequencing using the Illumina Paired-End Module. Paired-end sequence reads of 101 bases were generated using the Genome Analyzer IIX with v5 SBS reagent kits, as described in the Illumina Genome Analyzer operating manual. Data were processed using Real Time Analysis (RTA) v1.6.47.1. We generated an average of 17Gb PF (pass filter) data for each sample. A total of 93% of PF reads had a raw read accuracy of ≥Q30.

3.3 Whole exome sequencing

3.3.1 Description of Exome consensus

Authors: Jin Yu, Laura Clarke, Gabor Marth

For the Phase 1 Exomes, three different versions of capture platforms were used. The BCM-HGSC and BGI used the 'SeqCap EZ Human Exome Library' (v2.0 and v1.0 respectively) from Nimblegen, whereas the BI and WUGSC used the 'SureSelect All Exon V2 Target Enrichment' kit from Agilent. We first defined a consensus capture target region list by intersecting all the target design files (the .bed files) with the NCBI CCDS database, and then added 50 bp at either side of each consensus target. We used this extended target regions list for variant calling.

In the subsequent QC exercise of exome calls, we discovered that we had previously missed one of the three target design files from the intersection described above. Therefore, we had used a more extensive exome target set in our variant analyses. We updated the consensus target file by intersecting it with the target design file we previously missed and used this corrected version in exome data analyzing. Although the Exome variant calls were made from the more extensive targets, our comparison between SNPs found in the low coverage and the exome data (Figure 14) contains the correct intersection accounting for all three design files. As we always used a superset of the consensus target regions in exome variant calling, the shrinkage of the corrected version of consensus target only has a minimal effect to our results. The two sets of consensus targets files and README are hosted on the 1000 Genomes FTP[12].

3.3.2 Baylor College of Medicine – Human Genome Sequencing Center

Author: Donna Muzny

SOLiD Library Construction

DNA samples (5ug) were constructed into SOLiD precature libraries according to a modified version of the manufacturer's protocol (Applied Biosystems, Inc.). Briefly, the genomic DNA was sheared into fragments of approximately 120 base pairs with the Covaris S2 or E210 system as per manufacturer instruction (Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair (NEBNext End-Repair Module; Cat. No. E6050L) and A-tailing (NEBNext dA-Tailing Module; Cat. No. E6053L). The resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit (Cat. No. M2200L). Solid Phase Reversible Immobilization (SPRI) bead cleanup (Beckman Coulter Genomics, Inc.; Cat. No. A29152) was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR (LM-PCR) was performed using Platinum PCR Supermix HIFi (Invitrogen; Cat. No.12532-016) and 6 cycles of amplification. After bead purification, PCR products' quantification and their size distribution were analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. Primer sequences and a complete library construction protocol are available on the Baylor Human Genome Website[13].

SOLiD Exome Capture

The pre-capture libraries (2 ug) were hybridized in solution to NimbleGen EZ Exome 2.0 Solution Probes (~44 Mb of sequence targets from ~30K genes) according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture LM-PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over the background. Primer sequences and a complete capture protocol are available on the Baylor Human Genome Website[13].

SOLiD Sequencing Methods

The captured libraries were clonally amplified onto 1 um beads using emulsion PCR with a final library concentration of 0.70 to 0.85 pM. Emulsion PCR reactions were processed using either the Life Technologies Full-Scale 2 ePCR reaction protocol or a modified version of the Macro-Scale 8 ePCR reaction protocol. As specified by the vendor, the full-scale reactions used the IKA Ultra-Turrax to generate the emulsions followed by amplification with standard thermal cycling methods. The 8X bulk emulsions were generated using a Servodyne Electronic Mixer (Cole-Parmer, EW-50008-30, EW-50008-00) at a speed of 780 rpm for 20 min. The 8X bulk reaction was then amplified in a sealable bag using a Hydrocycler (K-Biosciences, HC-16) with the following cycling conditions: denature for 10 min at 95^{o}C, followed by 40 cycles of 1 min at 95^{o}C, 2 min at 62^{o}C and 2 min at 72^{o}C with a final extension of 10 min at 72^{o}C. Beads were recovered by centrifugation with 2-butanol and 50 ml conical centrifuge tubes and then enriched and 3' modified according to the Life Technologies Macro-Scale 8 ePCR reaction protocol.

The 3' modified template positive beads were deposited onto XD sequencing slides, targeting approximately 300 K beads/panel and sequenced using SOLiD V4 ToP reagents. Both barcode fragment and paired end sequencing methods were used in this project. For barcoded methods, capture libraries were individually captured and then pooled in sets of 4 samples after post-capture amplification. The 4 sample barcode library pools were sequenced with SOLiD Barcode Fragment Sequencing Kits (Life Technologies, 4452697). Here the first 5 bp barcode read is utilized to de-convolute the individual capture libraries followed by a 50 bp forward read. Individual capture libraries were sequenced with SOLiD Paired End Sequencing Kits (Life Technologies, 4459179) using a 35 bp reverse read followed by a 50 bp forward read. Base and quality calling for the SOLiD data was performed on-instrument using standard vendor software and settings.

3.3.3 Broad Institute

Author: Namrata Gupta

Whole exome library preparation was conducted using the same procedure described in the Low Coverage sequencing section. In place of size selection, in-solution hybrid selection was performed as described by Fisher *et al*[9]. Sequencing procedures were the same as described for the Low Coverage sequencing methods, with the exception that 76 bp paired-end reads were used.

3.3.4 Washington University

Author: Elaine Mardis

The Washington University Genome Institute utilized genomic DNA for exome sequencing that was provided by the Coriell Institute for Medical Research. Illumina libraries were constructed according to the manufacturer's recommendations (Illumina Inc, San Diego, CA), with the following exceptions: 1) 1 ug of DNA was sheared using a Covaris S220 DNA Sonicator (Covaris, INC. Woburn, MA) to a size range between 200-400 bp. 2) Four PCR reactions were amplified for 8 cycles to enrich for proper adaptor ligated fragments. 3) A Solid Phase Reversible Immobilization (SPRI) bead cleanup procedure was conducted to select size fractions between 300 and 500 bp. Hybridizations were performed utilizing the Agilent SureSelect Human All Exon v.2 kit. qPCR was used to determine the quantity of captured library necessary for loading. Two lanes of 2 X 101 paired-end reads on an Illumina GAIIx were generated in order to produce greater than 10 Gbp of sequence per sample. A minimum coverage of 70% of the targeted region at 20X depth was used to determine a passing sample.

3.4 OMNI genotyping

Authors: George Grant, Wendy Brodeur, Diane Gage, and Andrew Crenshaw

DNA samples were sent to the Broad Institute Genetic Analysis Platform for genotyping. Initially all samples were typed using a Sequenom MassArray SNP Genotyping panel of 23 SNPs and one gender determining assay to establish a genetic fingerprint. After gender concordance was verified the samples are then placed on 96-well plates using the Illumina HumanOmni2.5-Quad v1-0 B_SNP array. Omni genotypes were called using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 using the default cluster file HumanOmni2.5-4v1-Multi_B.egt. Called genotypes were run through a standard QC pipeline and only samples passing a call rate threshold of 97% and passing genetic fingerprint and gender concordance were passed. The Broad Institute did not filter any SNPs based on of technical quality control metrics. Only samples passing an overall call rate of 97% criteria and standard identity check were released.

3.5 Lane level identity checks

Authors: Laura Clarke, Xiangqun Zheng-Bradley, Richard E. Smith, and Petr Danecek

Each run was subsampled (typically 250-500Mb) and aligned to the reference genome. GLFtools[14] checkGenotype was then used to calculate genotype likelihoods for all the sites and these are compared to the known genotypes of all the samples. A total log likelihood for each possible sample is calculated and scaled according to the number of sites available for each possible sample. The ratio of the second most likely sample to the most likely sample is calculated and provided it is greater than 1.2 then the most likely sample is considered to be correct. If the most likely sample does not match expectations, or the ratio is less than 1.2, then the lane was withdrawn from further analysis and the production center responsible informed for further investigation. Many sample swaps could be resolved using this process, with the lane reassigned and reintroduced into the analysis process.

3.6 Post-alignment identity and contamination checks

Authors: Laura Clarke, Xiangqun Zheng-Bradley, Richard E. Smith, and Hyun Min Kang

The Project also performed sample identify and contamination checks on the alignment files using the VerifyBamID program[15]. The algorithm establishes whether the aligned reads match the genotypes from the given individual or another individual. The algorithm also can identify if an alignment is contaminated with non-sample DNA. For each aligned base that overlaps a known genotype, the probability that it was derived from a given individual is calculated. The program finds the individual whose genotypes best match the genotypes predicted by the alignment. If this does not match the expected sample the BAM file is further assessed to establish if this is problem is from just one of the runs contributing to the alignment or found in all of them. Any contaminated alignments are withdrawn from the analysis process and the production sequencing center notified to further investigate. As with the sequence level sample QC, if a run can subsequently be associated with the correct individual it was reinstated in the Project.

3.7 Analysis of cryptic relatedness and other sample identity checks

Author: James Nemesh

We analyzed genome-wide SNP data (generated using the Illumina Omni 2.5 array) to evaluate genetic relatedness of the samples eligible for sequencing by the 1000 Genomes Project. Our analysis used Plink[16] to generate an estimate of %IBD1 and %IBD2 (identity by descent) for each pair of individuals in each population (using the --genome command in Plink). These results were then categorized into different types of relationships using custom software written in R.

Three classes of relationships were categorized. "Parent-Child" relationships were defined as individuals sharing an entire haploid genome (criterion: 100% of their genomes IBD1.) "Sibling" relationships were defined by individuals sharing 25% of their genomes IBD2 and 50% IBD1. "Second-order" relationships refer to a class of relationships including uncle-niece, grandparent-grandchild, and half sibling, all of which involve an expected 50% IBD1 and 50% IBD0. IBD was calculated for each pair of individuals in a population, and the results were clustered by their IBD1 and IBD2 into these classes.

Any discovered cryptic relationships were validated by checking the expected IBD levels for the indirect, derived relationships implied by the cryptic relationships. For example, if two individuals in separate trios are found to be siblings, each sibling should have an avuncular relationship to the other sibling's child. The inferred relationships are shown in Table S10.

We conducted further analysis to detect potential sample purity or identity issues in each population. The problematic signal involves a modest level of predicted IBD1 "relationship" between one sample and many other samples in the population. This can arise when a DNA sample or cell line is contaminated with DNA or cells from another individual. In both of the cases we identified (NA20760 in TSI and NA20278 in ASW) the samples exhibited both a low level of IBD to many samples in the population, as well as having the highest heterozygosity in their respective populations.

4 Data processing

4.1 Low coverage Illumina and 454 processing

Authors: Shane McCarthy and Sendu Bala

Low coverage Illumina data was aligned to the reference using bwa v0.5.5[17]. First indexing the genome reference sequence was indexed using the command "bwa index -a bwtsw". To align, the command "bwa aln -q 15" was used to find suffix array coordinates of good hits for each individual read. A chromosomal coordinate sorted BAM file was then generated using the bwa sampe option for paired-end reads or the bwa samse option for unpaired reads.

Low coverage LS454 data was aligned to the reference using ssaha v2.5[18] by first precomputing the hash index using the command "ssaha2Build -skip 3 -save *$ref $ref*", where *$ref* is the reference fasta file. Reads were filtered to remove those shorter than 30 bp, then mapped independently using "ssaha2 -disk -454 -output cigar -diff 10 -save *$ref*". The top 10 hits for each read were recorded. The cigar output was then converted to BAM format, taking into account whether the reads were paired and the expected library insert size. For paired reads, 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.

The lane-level alignment BAMs were further processed to increase the quality and speed of subsequent SNP calling using tools from GATK[19,20] and samtools[21]. Reads underwent local realignment around known indels from the 1000 Genomes Pilot[6] using the GATK IndelRealigner command. Next, mate information in the resulting BAMs was fixed and coordinate sorted using the Picard package FixMateInformation command[8]. Read qualities were then recalibrated using the GATK TableRecalibration package, masking SNPs from dbSNP release 129. Finally, the command "samtools calmd –r" was used to introduce BQ tags[22] which could be used during SNP calling.

The processed BAMs were merged to create the release BAM files available for download. This process began by removing extraneous tags (OQ, XM, XG, and XO) to reduce total file size by around 30%. Next, lanes from the same library were merged using the Picard MergeSamFiles command, with PCR duplicates subsequently marked in these library-level BAMs (Picard MarkDuplicates). Library-level BAMs were merged (Picard MergeSamFiles) to the platform level, to produce a single BAM file for each of the sequencing platforms used to sequence each sample. Finally, the platform level BAMs were split into two separate BAM files - one containing all reads that mapped to the reference and one containing reads which did not map.

4.2 Low coverage SOLiD read mapping

Author: David Craig

SOLID fastq files were obtained from 1000 Genomes' DCC server based on the 2010/11/23 sequence index. Mate-pair 25mer reads and single-ended were not included in Phase 1 release as these samples were already sequenced on Illumina and/or 454 with longer paired reads. Reads were aligned to GRCh37 using BFAST[23] version 0.64e with the following settings: maxKeyMatches=8, maxNumMatches=384, queueLength=25000, local align offset=20, minMappingQuality=10, minNormalizedScore=36, and algorithm=3. SRR/ERR level outputs were realigned with GATK's indel realigner tool (V:1.0.04418)[20], followed by FixMateInformation within Picard[8]. Following realignment, lane level BAM files were recalibrated with GATK (V:1.0.4705). After recalibration BAM files had certain fields removed to only include RG, X1, NM, XT, MD, CS, CQ in order to reduce file size. These BAMs were then ready to be merged to sample level. After merging they went through Picard's mark duplicates tool. All BAMs were split by chromosome. Finally BAI, BAS, and md5sum files were generated for each BAM before they were transferred back to DCC.

4.3 Exome Illumina read mapping

Authors: Alistair Ward, Wan-Ping Lee, and Gabor Marth

Exome Illumina data was mapped at Boston College using the Mosaik pipeline described in Section 4.5 below.

4.4 Exome SOLiD read mapping

*Authors: Jeffrey Reid, Christie L. Kovar, and Fuli Yu**

** Corresponding Author*

Read and quality data were copied into our data-center where individual sequence events were split into 10M read bundles to undergo preliminary mapping using BFAST[23]. After read group bundles were mapped, their results were merged back into a single sequence-event-level BAM, and where necessary, these BAMs were merged into a sample-level BAM using Picard[8], and duplicate reads were marked at the library level using SAMtools[21]. Alignment metrics and uniqueness were evaluated to confirm that the sequencing performed as expected. To gauge the overall performance of the capture process, sample-level BAMs were also subjected to a capture analysis QC pipeline to obtain additional metrics such as the proportion of the aligned reads that mapped to the targeted region and the proportion of targeted bases at various coverage levels. Samples that met a minimum of 70% of the targeted bases at 20X or greater coverage were submitted for subsequent analysis and QC. In addition, sample concordance analysis was also performed by comparing SNP array genotypes to the sequencing data to confirm sample identity and evaluate contamination.

4.5 MOSAIK low coverage and exome read mappings

Authors: Gabor Marth, Chunlin Xiao, Erik Garrison, Wan-Ping Lee, Stephen Sherry and Alistair Ward

Phase I data was mapped using Mosaik[24] read mapping software collaboratively between Boston College (BC) and the National Center for Biotechnology Information (NCBI). A hash-seeded Smith-Waterman algorithm is utilized for all sequencing technologies in the project (Illumina, LS454 and SOLiD). The input parameters and specifics of the alignment pipeline were technology, read and fragment length dependent. The Mosaik aligner rescues unmapped or multiply-mapped read fragments by searching for mapping locations that meet library-specific fragment length and read orientation criteria. The identification of properly mapped pairs was performed using BCs BamTools[25] software. Duplicate marking was performed using Picard[8] for all Illumina and SOLiD data, and BCMMarkDupes was used for LS454 data. GATK[19,20] was used for base quality recalibration. All of the low coverage data were mapped at the NCBI and all of the exome data (including the official Illumina exome BAMs) were mapped at BC.

5 Variant calling

5.1 Low coverage and Exome SNP calling: Broad

*Authors: Guillermo del Angel, Ryan Poplin, Mark DePristo, and Eric Banks**

** Corresponding Author*

Low coverage SNP calling was performed on the project official low-coverage BAM files. Exome SNP calling was carried out using Illumina data only on BAM files produced locally at the Broad Institute using an equivalent pipeline to the official low coverage BAMs. These BAMs are available in a separate location on the project FTP site[26].

The Broad Institute produced a SNP callset for both the low coverage and exome samples using the GATK's Unified Genotyper. This multiple-sample, technology-aware SNP and indel caller uses a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of N samples, emitting an accurate posterior probability of there being a segregating variant allele at each locus as well as for the genotype of each sample. Mathematical details are given in DePristo *et al*[19].

Given a set of putative variants along with SNP error covariate annotations, variant quality score recalibration employs a variational Bayes Gaussian mixture model to estimate the probability that each variant is a true polymorphism in the samples rather than a sequencer, alignment or data processing artifact. The set of variants is treated as an *n*-dimensional point cloud in which each variant is positioned by its covariate annotation vector. A mixture of Gaussians is fit to a set of likely true variants. Here, we used the variants already present in HapMap 3 as well as those variants that were found to be polymorphic by the Omni chip were used as 'true' variants. Following training, this mixture model is used to estimate the probability of each variant call being true, capturing the intuition that variants with similar characteristics to previously known variants are likely to be real, whereas those with unusual characteristics are more likely to be machine or data processing artifacts.

The following error covariate statistics are calculated on a per-site basis and are used by the variant quality score recalibrator to model error.

- **QualByDepth**. The variant quality score (the confidence assigned by the unified genotyper in the site being a variant site) divided by the number of reads in the pileup. This statistic captures the intuition that as sequencing depth increases the confidence in the site should also increase if it is a real variant.
- **DepthOfCoverage**. The number of passing reads which cover this site.
- **HaplotypeScore**. A measure for how well the data from a 10 base window around the SNP can be explained by at most two haplotypes. In the case of mismapped reads, the pattern of mismatches around the SNP would seem to imply many more than two haplotypes and is indicative of error.
- **MappingQualityRankSum**. A Wilcoxon rank sum test that tests the hypothesis that the reads carrying the alternate base have a consistently lower mapping quality than the reads with the reference base.
- **ReadPositionRankSum**. A Wilcoxon rank sum test which tests the hypothesis that the alternate base is consistently found more often at the beginning or ending of the read instead of randomly distributed throughout. A bias would indicate that the reads are mismapped.
- **FisherStrand**. The p-value from a Fisher's exact test of the strandedness (positive or negative) of reads which hold the alternate allele versus those that hold the reference allele.
- **RMSMappingQuality.** The root mean square of the mapping quality of all reads covering this site.
- **InbreedingCoefficient.** The population genetics F-statistic. The degree of reduction or excess of heterozygosity when compared to the Hardy-Weinberg expectation.

5.2 Low coverage SNP calling: Baylor College of Medicine HGSC

*Authors: Yi Wang, James Lu, Fuli Yu**

** Corresponding Author*

The Phase 1 Low coverage data presented a challenge for reliable identification of SNP sites and accurate genotyping due to heterogeneity in sequencing technologies and the alignment methods. At the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC), we developed the SNPTools integrative pipeline[27] for the purposes of variant calling, which achieved high quality for (1) variant site discovery, (2) genotype likelihood estimation, and (3) genotype/haplotype inference via imputation[28]. It has demonstrated especially high performance when dealing with the Phase 1 low coverage data that was collected from heterogeneous platforms (Illumina, SOLiD, and Roche 454).

SNPTools applies a variance ratio statistic[28] to discover SNP sites. This statistic compares the difference in variation contributed from the variant read coverage (in case that they are true positives), and the variation that would be due to sequencing and mapping errors (in case that they are false positives). The larger the variance ratio statistic is, the more likely the site is a true polymorphic site. We identified 34.14 million SNPs from the Phase I samples, the overall Ti/Tv rate was 2.14. The novel sites were ~88.2% of our discoveries, and their Ti/Tv was 2.13, indicating high quality.

The next stage is to estimate the genotype likelihoods by clustering all candidate sites within each particular BAM file to overcome data heterogeneity using a binomial mixture model – this is named as the 'BAM-specific Binomial Mixture Model' (BBMM). BBMM normalizes the heterogeneity within each sample BAM by estimating BAM/sample specific parameters. It takes into consideration all the variant sites identified from the sample collection (30-40 million sites in ~1000 individuals), and clusters across sites in one particular BAM to estimate the genotype likelihoods. By clustering the scaled read coverage across millions of sites, it substantially reduces the variance and improves the accuracy in the genotype likelihood estimation. The BBMM can effectively overcome the heterogeneity in the low coverage data. The genotype likelihoods are passed to an imputation engine to refine the individual genotypes and produce phased haplotypes.

5.3 Low coverage SNP calling: University of Michigan

Authors: Hyun Min Kang, Mary Kate Trost, and Goncalo Abecasis

The UMAKE SNP calling pipeline[29] was used to produce low coverage SNP calls contributed by University of Michigan. The pipeline first computes genotype likelihood for each platform using the default samtools genotype likelihood model, after adjusting by per-base alignment quality (BAQ)[22]. Genotype likelihoods are merged across platforms if needed. This strategy effectively assumes dependency between base calling errors within a platform, but no dependency across platforms.

To detect polymorphic sites, we used Brent's algorithm[30] to obtain maximum likelihood estimates of allele frequency at each locus. We compared *Likelihood[no-variant]* to *Likelihood[variant]* under uniform prior between each 3 possible polymorphisms. Sites were considered as potentially polymorphic when the posterior probability of a variant call was ~0.70 (corresponding to a phred scaled quality score of 5), with neutral allele frequency spectrum under a constant population size at average mutation rate of 0.001. When calling variants on the X chromosome, the males are modeled as haploids except for the pseudo-autosomal regions (PAR) and females are modeled as diploids.

The candidate variant sites were filtered based on multiple per-site feature statistics, including (1) expected fraction of reference base at heterozygous allele (allele balance) (2) average depth across the samples (3) Pearson's correlation coefficient and z-score between strand and allele (strand bias metric) (4) correlation between machine cycle and allele (cycle bias), and (5) distance to nearby 1000 Genomes pilot 1 indels. For each of these features, a manually chosen threshold based on the empirical distribution of feature statistics, transition to transversion ratio, and overlap with HapMap SNPs was determined and filtered out the variants beyond the threshold. Each feature was independently considered during the filtering, and the sites that did not pass at least one criterion were filtered out from the call set.

From the procedure described above a total of 34,515,663 SNPs were identified. The transition to transversion ratio (Ts/Tv) for the variants overlapping with dbSNP build 129 was 2.14, and 2.16 for the rest of the variants. 98.63% of HapMap3 SNPs were rediscovered in this call set.

5.4 Low coverage SNP and indel calling: Sanger

Author: Petr Danecek

The Sanger low coverage SNP and indel calls were made by samtools and bcftools version 0.1.17 (r973:277)[21]. Samtools was used to generate all-site all-sample BCF files (samtools mpileup -C50 -m2 -F0.0005 -d 10000 –P ILLUMINA) with bcftools subsequently used to call variants (bcftools view -p 0.99 -bvcgs). Calling was conducted separately within the four continental groups, AMR, AFR, ASN, and EUR. The calls were then merged and filtered (StrandBias 1e-5; EndDistBias 1e-7; MaxDP 10000; MinDP 2; Qual 3; SnpCluster 5,10; MinAltBases 2; MinMQ 10; SnpGap 3). The pseudoautosomal regions on chrX (60001-2699520 and 154931044-155270560) were treated as diploid in male samples.

5.5 Low coverage SNP calling: NCBI

Authors: Chunlin Xiao, Tom Blackwell, Alistair Ward, Erik Garrison, Wan-Ping Lee, Hyun Min Kang, Mary Kate Trost, Gabor Marth, Goncalo Abecasis, and Stephen Sherry

The NCBI used a consensus calling strategy to generate a high-quality set of SNP calls. The strategy was based on the Boston College's freeBayes (version 0.4.2) and the UM's glfMultiples (version 2010-06-16) from a pool of 1094 samples, including 946 Illumina and 15 Roche454 BAM files generated with the Mosaik aligner, and 142 SOLiD BAMs generated with the Bfast aligner. The two SNP callers were used to generate two independent raw callsets with default settings. Then two raw callsets were intersected to create a consensus SNP callset to maximize confidence in the SNPs. The NCBI discovered 30,686,612 SNPs in the autosome with a transition to transversion ratio (Ts/Tv) of 2.29. 25.16% of the sites were previously known according to dbSNP Build129, and 98.4% of HapMap3 SNPs were rediscovered by this call set.

5.6 Exome SNP calling: Baylor College of Medicine HGSC

*Authors: Jin Yu, Danny Challis, Uday Evani, Fuli Yu**

** Corresponding Author*

For the Phase 1 Exomes, two different capture platforms were applied (see above). The BCM-HGSC and BGI used the SeqCap EZ Human Exome Library (v2.0 and v1.0 respectively) from Nimblegen, whereas the BI and WUGSC used SureSelect All Exon V2 Target Enrichment kit from Agilent. We first defined a consensus capture target region[31] by intersecting the two different target design files (the .bed files) with the NCBI CCDS database.

SNPs and indels were called using the Atlas2 Suite[32,33]. Atlas2 has capabilities for calling variants in high coverage next-generation sequencing data on multiple sequencing platforms (i.e. Illumina, Roche 454) for both SNPs and short-range (within tens of bps) indels. The Atlas2 Suite makes use of logistic regression models trained on whole exome capture sequencing (WECS) data to identify SNP and indel sites with high sensitivity and specificity, and subsequently produce accurate genotypes.

5.7 Exome SNP calling: University of Michigan

Authors: Hyun Min Kang, Goo Jun, Mary Kate Trost, and Goncalo Abecasis

The exome SNP calls were produced using the UMAKE SNP calling pipeline[29] consistent with low coverage calling with the following key differences. First, two call sets were generated separately for the Illumina and SOLiD platform. Second, SNPs were called only within 50 bp from the consensus target region. Third, the reads with mapping quality less than 20 were removed prior to calculating the genotype likelihoods. Fourth, the minimum posterior probability of a variant call was set to 0.90 (corresponding to a phred scale quality score 10). Finally, for the SOLiD exome sequence data where the base quality of aligned sequence reads were not empirically calibrated, we recalibrated the base quality using the GATK software (version 6128)[20] locally at Michigan with default parameters under the color space.

The candidate variant sites were filtered based on the per-site feature statistics used for the low coverage SNP calling, but the threshold for each feature was chosen differently based on the empirical distribution of the statistics. The thresholds were determined separately for the Illumina and the SOLiD platform.

Across the 822 samples sequenced on the Illumina platform, a total of 628,533 SNPs were identified, and 352,702 (56.1%) of them reside within the target region. The Ts/Tv for the dbSNP build 129 SNPs was 2.96 and 2.80 for the rest of the variants. 569,966 SNPs (90.7%) were included in the consensus exome SNPs. For the 306 samples sequenced in the SOLiD platform, 342,756 SNPs were identified, with 198,069 (57.8%) on target. The Ts/Tv ratios were 3.08 and 2.74 for SNPs within and outside of dbSNP129. A total of 324,730 (94.7%) of these SNPs were included in the consensus SNPs.

5.8 Exome SNP calling: Weill Cornell Medical College

Author: Juan Rodriguez-Flores

Genotypes for SNPs were called in 822 exomes sequenced on the Illumina platform and mapped using Mosaik. Genotyping was conducted in two phases: site discovery and genotyping. The site discovery phase involved identifying all possible SNPs in the 822 exomes, genotyping each exome separately and merging the sites discovered with a summary by population. The genotyping phase involved population-based genotyping at variant sites identified in the discovery phase. For each population (GBR, TSI, FIN, CEU, CHS, CHB, JPT, YRI, LWK, ASW, MXL, PUR, CLM), genotypes were called for all individuals of the population simultaneously, limiting the calling to sites identified in the discovery phase. The site list for genotyping was a consensus site list that excluded low-quality SNPs from the discovery phase that were identified by a support vector machine, and added high-confidence SNPs identified by other call sets.

The discovery site list included all autosomal SNPs identified by SAMTools[21] version 0.1.17 with SNP quality > 100 in one or more of the 822 exomes, limited to bases with quality > 17. In order to combine the individual calls, for each variant site the alternate allele was compared and verified to be consistent across samples. In cases where multiple alternate alleles were observed, the site was excluded. The depth and alternate allele count for each site was summed and included in the INFO column, with the depth limited to bases of quality > 17 summed across 822 exomes. The reported QUAL score is the lowest reported quality for an individual exome. A population-specific INFO tag was included, this lists for each SNP three-letter population codes where the SNP was observed in one or more exomes. Over 53% of SNPs were observed in only one population, with 5.4% observed in all 13 populations at least once. Called sites were limited to sites in the consensus target list +/- 50bp.

The WCMC site list was compared to call sets from Boston College (BC), Baylor (BCM), University of Michigan (UMich) generated from the same set of 822 of Illumina reads mapped using Mosaik. Concordance with other site lists was >98% when compared to BCM and UMich calls, lower (61.6%) when compared to BC. The total site list included 482,272 SNPs, with a Ts:Tv ratio of 2.94. The overlap with databases of variants includes 38.6% in dbSNP version 132, 6.3% in HapMap 3, and 67.2% in the consensus site list based on low coverage whole-genome sequencing. A fifth call set based on BWA alignments generated by the Broad was then compared to the four call sets based on Mosaik alignments. All five call sets included autosomal SNPs, and only the WCMC call set excluded sites with multiple alternate alleles. The Broad call set included indels. Both the Broad and UMich call sets include X and Y chromosome sites. The Ts:Tv ratio for the WCMC call set was within the range of other call sets for all SNPs, exonic SNPS, nonsynonymous SNPs and SNPs outside RefSeq transcripts, based on functional classification in RefSeq transcripts using ANNOVAR[34].

In the genotyping phase, the exomes were genotyped at all variant sites identified in the discovery phase, minus sites filtered out by the SVM filter and with the addition of high-confidence sites identified by the SVM filter in another call set (BC, UMich, BCM, or Broad). The SVM consensus site list includes 597,687 SNPs. No minimum quality score filtering was applied to this call set, and all observed alternate alleles were kept. The combined call set summary VCF file includes all alternate alleles observed (in order of decreasing frequency), the minimum quality score for a population where the variant was observed, and the combined depth and allele count across all populations. Population-specific information in the VCF file includes the number of populations where the SNP was observed (NPOP), and the alternate allele count observed in each population (POPAC). For sites where multiple alternate alleles were observed, the alternate alleles in each population are listed (POPALT). The Ts:Tv ratio for the major alternate allele was 2.8 (436,186 transitions and 155,281 transversions), excluding 6,220 sites in the SVM consensus site list where no SNPs were observed in this call set.

5.9 Variant calling from MOSAIK alignments

Authors: Erik Garrison, Alistair Ward, and Gabor Marth

SNP, MNP and indel calls were generated for the low coverage and the exome data using Mosaik alignments and BCs freebayes[35] Bayesian variant calling software. The variant calling pipeline included a left-realignment step (bamleftalign[35]) to ensure that all indels were left-aligned in order to be consistent with the project conventions. In some situations, local misalignment of larger indels results in spurious, artifactual mismatches or gaps. BCs ogap[36] software was used to realign all reads with embedded gaps using alignment parameters designed to replace multiple putative in-phase indels and SNPs with lesser number of larger events. This process does not introduce new mismatches or gaps into the alignments. Finally the BAQ model implemented in samtools[21,22] (samtools calmd) was applied to the reads to incorporate local alignment quality into base quality. Following these preprocessing steps, variants were called by assessing the BAM files from all populations simultaneously and all bi- and multi-allelic variants were reported in the VCF format[37]. Genotypes and genotype likelihoods were also generated for each sample at each variant site.

In order to consider a variant allele, low coverage data required a minimum of two observations of the alternate allele. Due to increased coverage in the exome data, this requirement was increased to a minimum of five reads supporting an alternate allele in order to consider the variant. The Bayesian model implemented in freebayes[35] establishes the posterior probability that a given locus is polymorphic in the samples under analysis given a neutral model for allelic diffusion and differentiation. The prior probability model combines an estimate of the probability of sampling a given set of allele frequencies provided an expected pairwise heterozygosity rate (Ewens Sampling Formula) with the discrete sampling probability of the set of genotypes given the allele frequencies, which effectively incorporates the neutral expectations of genotype frequencies under Hardy-Weinberg equilibrium. The resulting variants were filtered on estimated posterior probability of polymorphism and alternate allele balance among heterozygotes using BCs vcffilter (part of the vcflib[38] package) to remove low quality SNPs.

BC called 33,324,407 SNPs in the autosomes of the 1,094 samples. 23.8% of these were known sites (contained in dbSNP). The TsTv ratio for these sites was 2.12 (2.1 for novel sites and 2.17 for known). The Illumina exome data (822 samples) yielded 344,781 SNPs with a TsTv ratio of 3.18 (3.09 for the novel sites and 3.52 for the known sites. 22.1% of the exome sites were previously known). The SOLiD exome data (306 samples) yielded 176,637 SNPs with a TsTv ratio of 3.34 (3.22 for novel sites and 3.58 for known sites. 36% of the sites were previously known).

BCs vcfCTools[39] software was used to perform set analysis (union, unique and intersections) between the variant calls made at BC and those made by the Broad, Baylor, NCBI and Cornell.

5.10 Creation of low coverage SNP consensus

*Authors: Guillermo del Angel, Ryan Poplin, Mark DePristo, and Eric Banks**

** Corresponding Author*

The GATK[19,20] was used to generate a consensus project callset from the six individual sets submitted by the various centers. A high-level view of the process used to build the project consensus is as follows:

1) First pool together all SNP calls made by any center. For the Phase 1 calls this list was approximately 46.3 million SNPs.
2) Re-call at all SNP sites using the GATK's Unified Genotyper with project BAM files that in addition have been fully indel realigned at the population level. The calls were made by dividing the samples into nine overlapping analysis panels as follows:

- EUR = CEU + FIN + GBR + TSI + IBS
- EUR.admixed = CEU + FIN + GBR + TSI + IBS + MXL + CLM + PUR + ASW
- AFR = LWK + YRI + ASW
- AFR.admixed = LWK + YRI + ASW + CLM + PUR
- ASN = CHB + CHS + JPT
- ASN.admixed = CHB + CHS + JPT + MXL + CLM + PUR
- AMR = MXL + CLM + PUR
- AMR.admixed = MXL + CLM + PUR + ASW
- ALL populations

3) The Unified Genotyper additionally adds several important statistics, calculated on a per-site basis, which will be used by the variant quality score recalibrator. These statistics were explained in more detail above and are the same as those used to produce the Broad Institute SNP callset.
4) Apply the Variant Quality Score Recalibrator genome-wide to train a Gaussian mixture model over the same eight per-site error covariates listed in Section 5.1

(QualByDepth, DepthOfCoverage, HaplotypeScore, MappingQualityRankSum, ReadPositionRankSum, FisherStrand, RMSMappingQuality, and InbreedingCoefficient). Each input variant is assigned a VQSLOD score, which is the log odds ratio between the probability that the SNP is true or false given the model. In addition to using the error covariate statistics the model incorporates a prior probability of being a true variant which is based on the number of callsets that the original variant is found in. This captures the intuition that variants called independently by multiple callers are more likely to be real. The prior used here is Q<10X>, where X is number of callsets the variants was found in.
5) Partition the list of variants into those that are PASSing and those which are filtered out by choosing the VQSLOD value which gives 99.8% sensitivity to the accessible HapMap3 variants.

This procedure produces a high quality and statistically principled consensus set of sites.

5.11 Generation of consensus Exome SNP call set

Authors: Hyun Min Kang, Goo Jun, Mary Kate Trost, and Goncalo Abecasis

For each Illumina and SOLiD platform, a union exome call set (across 5 sets for the Illumina platform and across 3 for the SOLiD platform) was produced. Each union call set was filtered jointly by multiple criteria using Support Vector Machine (SVM) approach described as follows.

First, for each candidate variant site, per-site feature statistics including allele balance, strand bias, cycle bias, average depth, and inbreeding coefficient statistics were calculated from aligned sequence reads.

Second, a preliminary filter was applied for each feature separately based on a threshold manually determined from the empirical distribution of the feature statistics, similar to the filtering step in the Michigan Exome SNP calling. Then each site is annotated by each filtering criterion whether the site met (PASS) or did not meet (FAIL) the criterion. In addition, each individual call set was considered as an additional filter by annotating each as present (PASS) or absent (FAIL) in the individual call set. After this preliminary filtering step, each candidate variant site is annotated by multiple PASS or FAIL labels across multiple filtering criteria.

Third, we apply a SVM model by labeling sites filtered out by three or more criteria as negative labels, and considered sites overlapping with HapMap SNPs as positive labels. Each feature statistic was normalized using an inverse normal transformation, and a SVM model with Gaussian radial was applied with the assigned label to score each variant with a SVM score. Variants with positive SVM score were considered as consensus SNP.

In the SOLiD exome consensus call set, the singletons exclusively called by Baylor College had noticeably smaller overlap with low coverage SNPs with much lower deamination (G->A,C->T) to deamination (A->G,T->C) ratio compared to the other call sets. We refined the consensus call set by additionally filtering out the variants supported by less than 3 effective variant reads (see Baylor Exome SNP calling for details) from the variants exclusive to Baylor's call sets. As a result, 6,805 SNPs were additionally filtered out from the consensus call set.

A total of 597,695 SNPs and 356,114 SNPs passed the SVM consensus approach for the 822 Illumina samples and the 306 SOLiD samples, respectively. The Ts/Tv ratios for SNPs known and novel to dbSNP129 were 2.98 and 2.74 for the Illumina call set, and 3.09 and 2.91 for the SOLiD call set. The SVM consensus approach produced better quality metrics than the consensus call set by simple voting strategy. For example, when a 3-out-of-5 or 2-out-of-3 consensus approach was used for the Illumina and SOLiD platforms, respectively, slightly fewer SNPs passed the criteria than SVM consensus call set, but the overlaps with SNPs monomorphic in the Omni2.5 genotype platform increased by 25% and 36% for each platform.

5.12 Low coverage and exome indel calling: Broad

*Authors: Guillermo del Angel, Ryan Poplin, Mark DePristo, and Eric Banks**

** Corresponding Author*

As for SNP calling at the Broad, low coverage indel calling was performed on the project official low-coverage BAM files. However, exome indel calling was carried out using Illumina data only on BAM files produced locally at the Broad Institute using an equivalent pipeline to the official low coverage BAMs. These BAMs are available in a separate location on the project FTP site[26].

The Broad Institute produced an indel dataset consisting of genomic sites and genotype likelihoods for all low coverage samples, as well as for the 822 exome samples sequenced with Illumina technology. The production of both datasets followed the same procedure:

1. Realignment around Indels in reads
The purpose of this step is to create consensus indels in the reads so that base mismatches are minimized. Each sample was realigned independently, considering as candidate sites known indels, as well as sites where the read mapping software introduced insertions or deletions.

2. Choose candidate sites and alleles to genotype
For each indel present in reads after realignment, a simple counting algorithm was used to create candidate indel sites and alleles: if a candidate indel allele was present in at least 5 reads at a site, it would be passed over to the next step for genotyping, or otherwise it was excluded. At most only one alternate allele was output per site (i.e. only biallelic calls were produced). If a site had multiple alternate alleles present, the allele with highest count in reads was chosen.

3. Genotype candidate alleles on each sample

For each candidate site, the genotype likelihoods were computed for each sample. Likelihood computation involved forming 2 candidate haplotypes per site, one containing only the reference allele and the other containing the alternate allele, and then scoring each sample's read against each of these 2 haplotypes. The likelihood of each read scored against each haplotype was computed using a Pair Hidden Markov Model[40] using affine gap penalties.

4. Compute site qualities and attributes and post-filter variants

Based on the Genotype Likelihoods computed at the previous step, the Allele Frequency distribution was computed for each site. Only sites with a probability of being variant that exceeded 0.6 (Phred-scaled Q value 4.0) were kept.
Additionally, several site attributes were computed. In particular, the following attributes were computed in order to serve as statistics for subsequent filtering, with each attribute's computation and meaning being the same as in the SNP case:

- FisherStrand
- QualByDepth
- ReadPosRankSum
- InbreedingCoeff

Only variants whose attributes fell within empirically derived thresholds for each annotation were kept. After filtration, the resulting low coverage callset had 5,543,104 variants, out of which 1,651,867 were insertions, 3,578,869 were deletions and 312,368 were complex substitutions. The resulting exome callset had 11,240 insertions, 21,944 deletions and 1,723 complex substitutions.

Exome indels were called using the same statistical algorithm as described above but due to the limited number of indels in the exome callset machine learning of error modes was not possible. Consequently the following hard filters were applied:

- QualByDepth < 2.0
- ReadPosRankSum < -20.0
- InbreedingCoeff < -0.8
- FisherStrand > 200.0

5.13 Low coverage indel calling: Sanger

Author: Petr Danecek

Low coverage indel calls from the Sanger were called by samtools[21] using the procedure described in the SNP calling section (Section 5.4).

5.14 Low coverage indel calling: Dindel2

Author: Kees Albers

The Dindel2 algorithm realigns reads to candidate haplotypes using a probabilistic approach based on the read-haplotype alignment probabilistic model described in Albers *et al.*[41]. The main differences are that Dindel2 uses a multi-sample haplotype caller based on a model selection approach rather than a pooled Bayesian EM algorithm, and that it uses a banded gapped alignment of the read to two seed positions in the candidate alignment.

Individuals in the same population group were analyzed jointly. First, all indels identified by the read mapper (BWA) were extracted from the alignments. All of these indels were tested by realigning the reads against candidate haplotypes consisting of the reference haplotype and the alternative haplotype resulting from the indel. Second, all indels called in the first step were retested, however, allowing for at most two nearby SNPs, thus potentially realigning all reads to at most 8 candidate haplotypes consisting of candidate SNPs and one candidate indel. Third, all indels called in the second step were subjected to a post-analysis filter step. All indel calls satisfying at least one of the following criteria were filtered: Bayes factor for strand bias was >4.0, called from only one fragment, indel is in a homopolymer run longer than 10 nucleotides, more than 90% of reads have mapping quality below Q30, or more than 90% of the reads have the indel positioned in the first or last 10 bases.

5.15 Low coverage indel calling: Oxford

Author: Gerton Lunter

Platypus is a haplotype-based variant caller[42]. The program integrates the calling of SNP and indel variants of up to 50 bp, using a 3-step process. First, candidates for SNP and indel polymorphisms are generated usiong the input reads from all population samples and their alignment to the reference sequence. Second, haplotypes are generated from sets of these candidate variants restricted to small windows, and all reads are re-aligned to these haplotypes. Third, an EM algorithm estimates the frequencies of the haplotypes in the population, and determines which haplotypes are supported by the data; the set of haplotypes that have support determine the variants that are reported to be segregating in the population.

To remove poorly or ambiguously mapped reads, Platpyus requires a minimum mapping quality of 20 on the Phred scale. This filtering improves the robustness of calls and reduce the number of spurious candidates. In addition, duplicate reads are removed to reduce the impact of non-independent errors.

Variant candidates are considered by Platypus if they are seen at least twice. For SNPs, the variant base must be seen at least twice with base-quality exceeding 20. Indel candidates are left-normalised. Platypus then looks in small (~100-200 base) windows across the genome, and creates haplotype candidates, based on the list of variants in each window. Each haplotype may contain several variants. As the number of possible haplotypes is generally exponential in the number of candidate variants, the program adapts the window size and implements some heuristic filters to limit the number of haplotypes that are considered to 256.

An EM algorithm is used to infer the population frequency of each haplotype in the data provided. This algorithm, which includes priors for SNP and indels, and a model for genotype frequencies given the frequencies of variants, works by re-aligning all of the reads to each of the haplotypes, and computing a likelihood for each read given each possible diploid genotype. The algorithm used to calculate these genotype likelihoods includes a model for indel errors in Illumina reads, similar to the model used by Dindel[41]. Platypus uses the inferred frequencies and the likelihoods to compute a probability for each variant candidate segregating in the data. These probabilities are reported in the VCF output file.

Finally the variants are filtered to reduce the false-positive rate. First, variants are only called if they have a high enough posterior probability (Phred score exceeding 20). Additional filters are used to remove variants that are only supported by reads on the forward or reverse strand.

For the analysis of Phase 1 data, we considered both SNPs and indels during the calling process, but only indel calls were reported. The resulting VCF file contains 4,904,406 indel calls ranging from length 1 to 63 bp.

5.16 Creation of low coverage indel consensus

*Authors: Guillermo del Angel, Ryan Poplin, Mark DePristo, and Eric Banks**

** Corresponding Author*

The creation of a low-pass Indel consensus set was similar to the SNP consensus creation approach. First a union of all five input datasets was produced. In order to create this union, every genomic position in each input dataset was left aligned in order to guarantee that it had a consistent representation and that sites with differing coordinate positions but which encode the same alternate allele were combined correctly.

This union had a total of 30,720,770 indels. For each input site in this union, the following steps were performed:

a) Creation of genotype likelihoods and site annotation, following exactly the same procedure as outlined above for the creation of the SNP consensus set.
b) Training a VQSR model[19] based on the following statistics:
    - FisherStrand
    - QualByDepth
    - ReadPosRankSum
    - InbreedingCoeff
    - HaplotypeScore
c) The reference data used for training was the site list produced in Mills *et al.*[43], subsetted to sites that were seen in at least 2 centers and in at least 2 traces from Sanger sequencing. This training data set contained 832,595

sites, consisting of 345,859 insertions, 334,723 deletions and 152,013 complex events.

d) Cutting data set to keep only the best fits to the VQSR model. A cutting threshold was chosen to keep 95% of the training variants in the output result. This resulted in a data set containing 1,648,546 simple insertions, 2,343,430 deletions, and 1,515,693 complex or multi-allelic records.

Given the lower quality of multi-allelic indel records, it was decided that only the biallelic records would be kept and sent forward for haplotype integration with SNPs.

5.17 Structural variation: Deletions

Authors: Robert E. Handsaker, and Steven A. McCarroll*

** Corresponding Author*

The set of structural variants included in the integrated call set was limited to large (greater than 50bp) bi-allelic deletions. Site selection of these structural variants was done in three steps: First, a list of candidate sites was chosen by combining deletion calls from five deletion discovery algorithms (BreakDancer, CNVnator, Delly, Genome STRiP, and Pindel) plus the set of deletion calls from the 1000 Genomes pilot that had assembled breakpoints. Second, a subset of these candidate sites was selected for genotyping based on estimating the overall false discovery rate using the Omni 2.5 SNP array intensity data. Third, the set of sites contributed to the integrated call set was selected after genotyping based on (a) whether there was sufficient data available at the site to generate well-calibrated genotype likelihoods (b) removal of redundant overlapping calls of the same underlying polymorphism and (c) removal of sites that were classified as false discoveries based on the genotyping results.

Five deletion discovery algorithms were used to make independent calls of large deletions (longer than 50 base pairs). These were combined with a site list from the 1000 Genomes pilot consisting of 10,855 deletions that had assembled breakpoints in the pilot to yield a (potentially redundant) set of 113,649 candidate sites. The five computational methods used for selecting the list of candidate sites are described below.

5.17.1 BreakDancer (run at WTSI)

Authors: Klaudia Walter and Ken Chen

Deletion calls were made with BreakDancerMax1.1 for 929 samples[44], all paired-end sequenced on the Illumina platform and mapped with BWA[17]. Only paired-ends with mapping quality of at least 20 were considered. Insert size distributions were analyzed for chromosome 20 for each library separately to determine thresholds for each as upper cut-off in the BreakDancer config files. To accommodate the variety of insert size distributions, three different types of thresholds were calculated, (1) the drop in the density function for each insert size distribution, (2) the median plus four times the standard deviation, (3) the median plus five times the median absolute deviation (MAD), then the maximum of those three types was used. In a few cases when the median insert size was zero, the cut-off 1,000 was chosen, and if the third quantile of the insert distribution was zero, the cut-off 10,000 was chosen. The raw BreakDancer calls were filtered for deletion size (≤ 50 bp and > 1 Mb), for estimated copy number (< 0 and ≥ 2), for number of spanning read pairs (≥ 20), for regions around centromeres (+/- 1 kb), for regions around assembly gaps (+/- 50 bp) and for alpha satellite regions. Deletions were then merged across samples if there was a 50% reciprocal overlap with connected components. The merging process generates confidence intervals for the start and for the end position of the deletion that were used for further filtering, *i.e.* if the upper confidence limit for the end position was lower than the lower confidence limit for the start position, or if the confidence interval was larger than 10 kb. To improve the specificity of the call set, the deletion set of the 1000 Genomes Pilot Project was used as training set[6]. A likelihood ratio was computed using the attributes deletion size, BreakDancer score, number of samples, estimated copy number and number of libraries. The breakpoints were estimated by centering the deletion within the outer confidence limits and by using the deletion size estimate from BreakDancer.

5.17.2 CNVnator

Author: Alexej Abyzov

To call CNVs with CNVnator, data from all individuals within each population were pooled together. The data were processed with CNVnator software[45] (version 0.2.2) with standard settings and 100 bp and 50 bp bins. Additionally, CNVs were called with relaxed parameters to call for lower allele frequency CNVs. For each population, overlapping calls were merged by selecting the largest call of all overlapping ones. Merged calls were filtered. A CNV call passes the filter if: i) it does not overlap a gap in the reference genome for a deletion, and it is not within 0.5 Mb from a gap in the reference genome for a duplication; ii) a deletion must have at least two paired-end reads supporting the predictions (overlapping by 50% reciprocally) or read depth in the middle part (1 kb away from breakpoints) of the called region should be statistically different from average read depth. Statistical testing was done the same way as when calling CNV regions.

5.17.3 Delly

Authors: Tobias Rausch and Jan Korbel

DELLY[46] (version 0.0.1) integrates paired-end mapping with split-read refinement. The paired-end analysis step[47] relies on the identification of discordantly mapped paired-ends, which show an alignment distance (insert size) on the genome that deviates significantly from the expected distance. All discordantly mapped paired-ends are clustered and merged to estimate the putative start and end coordinate of the SV. Each paired-end SV interval is then screened for split-read support. All collected split-reads are grouped together and their consensus sequence is aligned to the reference to detect the SV at single-nucleotide resolution, including any microinsertion and microhomology present at the breakpoint.

5.17.4 Genome STRiP

Authors: Robert E. Handsaker, and Steven A. McCarroll*

** Corresponding Author*

We used the Genome STRiP algorithm[48] (Version v1.04.683) for large deletion discovery using information from read pairs with unexpected alignments and analysis of sequencing read depth. Low coverage Illumina sequencing data from 946 samples was analyzed together to perform discovery. We ran the Genome STRiP algorithm with default parameters plus two more stringent filters for this phase of the 1000 Genomes project:

a) (alpha satellite repeat) Sites were removed if at least 90% of the deletion site was annotated as alpha-satellite repeat (based on the UCSC hg19 RepeatMasker annotations).
b) (multiple read pairs per genome) Sites were required to have an average of at least 1.1 aberrantly aligned read pairs per genome having any observed aberrant read pairs at this site. This removed sites that were supported predominantly by observing only one aberrant read pair in each putative carrier individual (usually across only two or three putative carrier individuals).

5.17.5 Pindel

Author: Kai Ye

An improved version of Pindel[49] (version 0.2.0) was used to call insertions, deletions, inversions, and tandem duplications from Phase 1 Illumina low coverage sequence data. All samples were processed at the same time for joint variant discovery of all variant types simultaneously. All reads with suspicious mapping status such as soft-clip and >5% mismatches were subjected to Pindel re-alignment and up to 3% of the read length was allowed for mismatch. The new version of Pindel can find deletions, inversions and tandem duplications even in the presence of non-template bases inserted at the edges of the structural variation, and also reports the non-template bases. The variants are selected if larger than 100 bp, appear in more than 3 samples and with more than 10 supporting reads in total among all samples.

5.17.6 Merging candidate deletion calls to create sensitive set

Author: Marcin von Grotthuss

Calls from the five algorithms were merged along with sites derived from assembled breakpoints from the 1000 Genomes Pilot Phase[6] to create a merged set of 113,649 potential deletion sites.

First, we determined the confidence intervals of each computational call set by comparing deletion calls with calls from the 1000 Genomes Pilot Phase that had assembled breakpoints. We required 80% reciprocal overlap to match deletion calls with the pilot phase assembled deletions as a filter to avoid comparing calls that correspond to different deletions. Start and end position residuals were obtained from each matched call, resulting in the distributions of deviations from the actual event. The 95th percentile of the observed deviations was assigned as the confidence interval of each call in the deletion call set.

Next, the calls from 5 deletion call sets, plus the calls from the 1000 Genomes Pilot with assembled breakpoints, were merged using hierarchical clustering with complete linkage with a decreasing reciprocal overlap from 100% to 80%. We allowed deletion calls to be merged together only when the confidence intervals intersected.

If the predicted breakpoints of the merged calls were outside of the intersection of the confidence intervals, the midpoint of the intersection was assigned as the most likely breakpoint of the merged call, pending breakpoint assembly, and the innermost coordinates of the intersecting confidence intervals were assigned as the innermost merged confidence intervals (CIIPOS/CIIEND). If there were any predicted breakpoints within the intersection, then the midpoint between the predicted breakpoints was used as the most likely breakpoint. If deletion calls were merged with a call from the pilot set, the assembled breakpoints were used as breakpoints of the merged call, and the intersecting confidence intervals were set to zero. We did not allow calls from the pilot set to be merged together since they were assumed to correspond to different deletions. The outermost coordinates of the confidence intervals of the merged calls were used as the values of the CIPOS/CIEND intervals.

5.17.7 Assembly of deletion breakpoints

Author: Alexej Abyzov

Breakpoint assembly was attempted on the set of 113,649 merged candidate deletion calls. Assembly was done by first generating local assemblies of candidate contigs for the alternate allele using TIGRA-SV[50] and then aligning these contigs to the reference genome assembly using both CROSSMATCH[51] and AGE[52]. Breakpoints obtained from either aligner were used and in cases where the two aligners differed the AGE alignments were used.

Out of the 113,649 merged candidate deletions calls, unique breakpoints were assembled for 50,776 loci, of which 8,934 are in the set of genotyped large deletions.

5.17.8 Performing local assembly with TIGRA-SV

Author: Ken Chen

Breakpoint assembly was performed using TIGRA-SV (version v0.3.0) on a set of 113,649 merged candidate deletion calls. For each call, TIGRA-SV[53] first obtained reads surrounding the predicted breakpoints (+-500 bp) from the set of bam files that were predicted as deletion containing. It then ran a *de Bruijn* graphic assembly algorithm to decode the set of non-reference alleles that best explain the set of reads. 111,353 calls were successfully assembled (i.e., obtained at least one contig). An assembly score was calculated to summarize both the length of the contigs and the amount of reads that contributed to the results. The assembled contigs were aligned with CROSSMATCH[51] and AGE[52] as described below to yield breakpoints for each deletion call.

5.17.9 Deriving breakpoints from CROSSMATCH alignments

Author: Ken Chen

Contigs locally assembled with TIGRA-SV[50] were aligned using CROSSMATCH[51] (version 1.080721) against corresponding 1000 Genomes v37 reference sequences that span the putative deletions with 700 bp flanking sequence on either end. A deletion is called "validated" and is passed to the next stage if the associated pair-wise alignments indicate the existence of the same deletion as was predicted by the original callers. A deletion is not validated if the alignment was ambiguous, i.e., contain more than 2 high scoring pairs and have an assembly score < 200, or if the size differed by more than 50% from the expectation. In total, 61,265 deletions had some evidence of assembly support. 38,020 were called "validated" by CROSSMATCH.

For each validated deletion, its precise boundary as well as patterns of target site duplication and non-template insertion were deduced from the alignment and recorded in the VCF files.

5.17.10 Deriving breakpoints from AGE alignments

Author: Alexej Abyzov

Contigs locally assembled with TIGRA-SV[50] were aligned with AGE[52] (version 0.2) to target deletion regions extended by 1 kbp downstream and upstream. AGE was run with options '-indel –both' and the following scoring parameters: match=1, mismatch=-1, gap_open=-10, and gap_extend=-1. Typically each deletion region had few alternative contigs assembled. Each one was aligned to target region.

Each AGE alignment consists of 5' aligned sequence (left flank), excised region suggestive of an SV, and 3' aligned sequence (right flank). For each deletion breakpoints were assigned as coordinates of excised region from a contig alignment that satisfies all the following requirements: i) the contig is at least 100 bps in length; ii) at least 90% of contig bases are aligned; iii) alignment of the contig has excised region suggestive of a deletion compared to the reference genome; iv) length of each alignment flank is at least 30 bps (regions of sequence micro-identity around excised region are not included in the lengths calculation); v) contigs has no more than one alternative alignment of equal score; vi) average alignment sequence identity in right and left flanks should be at least 98%; vii) alignment sequence identity in each flank should be at least 97%; viii) coordinates of excised region and target region should overlap reciprocally by at least 50%; ix) each coordinate (i.e., start and end) of target region and excised region should not differ by more than 200 bps; x) for an alternative alignment the previous two requirements should also be satisfied. In case more than one contig satisfies the requirement then one of them was chosen randomly and its alignment was used as the inferred breakpoint.

5.17.11 Classification of SV formation mechanism by BreakSeq

Authors: Alexej Abyzov, Jasmine Mu, Robert E. Handsaker*

** Corresponding Author*

Using BreakSeq (version v1.3), SVs were classified according to their likely mechanism of formation[54]. In particular, SVs were classified into the following formation mechanisms: (1) non-allelic homologous recombination (NAHR); (2) non-homologous rearrangements (NHR), including non-homologous end-joining (NHEJ) or microhomology-mediated break-induced replication (MMBIR); (3) variable number of tandem repeats (VNTR); and (4) mobile element insertions (MEI). The fraction of deletions classified as NAHR and NHR are roughly consistent for deletions with lengths above 500 bp, the size range in which we had good power to genotype these events (Table S9).

5.17.12 Creation of the specific SV discovery set and FDR estimation

Author: Robert E. Handsaker

The set of 113,649 deletion sites from the sensitive SV discovery set were filtered prior to genotyping to create a more specific call set and ensure a low false discovery rate (FDR). This more-specific call set was constructed to have a FDR below 5%, based on the Omni 2.5 validation results, the only ones available at the time. Based on the estimated FDR of 1.5% for the Genome STRiP call set, all calls made by Genome STRiP were promoted to the specific call set. To this set, we added sites from any other caller with a rank-sum p-value ($p < 0.01$) based on the Omni 2.5 probe intensity. This yielded a more specific set of candidate deletion sites containing 23,592 sites and this set, by contruction, has a predicted FDR of less than 5%.

The Omni 2.5 results and subsequent validation experiments using PCR and Array CGH provide estimate of the FDR for the Genome STRiP calls ranging from at 1.5% - 4.2%. Assuming a FDR of 1% for the calls promoted based on Omni 2.5 rank-sum p-value ($p < 0.01$), we used weighted averaging to estimate the FDR for the 23,592 deletion sites at 1.4% - 3.7%, corresponding to the different FDR estimates for the Genome STRiP calls.

An alternative approach to estimating FDR of the specific SV discovery set is to utilize the PCR and Array CGH results, where the validation sites were selected independently of the construction of the specific call set. In total, validation was attempted on 3490 of the 23,592 sites (3404 by Array CGH, 185 by PCR, with 98 of these sites subject to both validation methods). Discarding sites with ambiguous or discordant validation results yields unambiguous validation results for 3415 sites, of which 3343 validated and 72 were invalidated, yielding an estimated FDR of 2.1%.

The full set of merged deletion sites and the more specific subset are available in a supplemental data file in VCF format. In this file, the FILTER column is used to indicate whether each site was included in the more specific deletion set. A FILTER value of NONVAL indicates sites that did not meet the criteria for inclusion in the more specific discovery set. A FILTER value of PASS indicates sites that were genotyped and included in the integrated call set. Other FILTER values indicates sites that were in the specific discovery set but were not included in the final list of genotyped sites integrated with SNPs and small indels because (a) they were not on the autosome or chromosome X (NONAUTX) (b) there was insufficient data to obtain accurate genotype likelihoods (NONGT) or (c) after obtaining genotype likelihoods, they were confidently non-polymorphic (NONVARIANT) or likely redundant calls at the same locus (DUPLICATE).

5.17.13 Structural variation genotyping

Authors: Robert E. Handsaker, and Steven A. McCarroll*

** Corresponding Author*

Genotyping was performed on the 23,592 candidate high-specificity deletion sites using Genome STRiP[48] (version v1.04.784) utilizing read depth and aberrant read pairs to generate genotype likelihoods for 946 samples using Illumina low coverage sequencing data. Split read alignments to the alternate alleles described by the assembled breakpoints were not used in calculating the genotype likelihoods due to the potential for local assembly errors to impact the genotype likelihoods.

The candidate sites were filtered after genotyping to eliminate poorly performing sites, sites that were confidently non-polymorphic and sites that appeared to be redundant calls of the same underlying polymorphism. The filters utilized both the posterior genotype likelihoods and estimated model parameters from the Genome STRiP genotyping model for read depth, which fits a Gaussian mixture model to the normalized read depth for each sample. The following post-genotyping site filters were applied:

a) (read depth cluster separation) We measured cluster separation by the mean (across all samples) of the distance between the expected locations of the copy number 1 and copy number 2 clusters divided by the square root of the mean of the variances. The cluster separation between the copy number 1 and copy number 2 clusters was required to be at least 2.0 (at least 2.5 for deletions larger than 100 kb).
b) (excessively low/high read depth) The estimated centers of each copy number cluster were required to be between 50% and 150% of the

expected read depth based on genome-wide average sequencing coverage.

c) (sufficient uniquely alignable sequence) Each site was required to have an “effective alignable length” of at least 100 bp, defined as the number of base positions where a 36 bp window centered over the base position is unique within the reference genome.
d) (redundant call removal) If two independently called sites had at least 50% reciprocal overlap and none of the genotyped samples had discordant genotypes where the joint likelihood of discordance for that sample was more than 99%, then the site with the lowest total posterior genotype likelihood was eliminated as being a redundant call of the same polymorphism.
e) (evidence of polymorphism) Sites were removed if the posterior genotype likelihoods were at least 95% confident homozygous reference for every sample.
f) (excess heterozygous genotypes) Sites were removed if the inbreeding coefficient calculated across all samples was < -0.15.

The final set of genotyped sites contained 14,422 deletions on the autosome and X, each with genotype likelihoods for 946 samples, which were then integrated with the genotypes for SNPs and short indels.

5.17.14 Structural variation genotyping on chromosome Y

Authors: Robert E. Handsaker[*]*, and Steven A. McCarroll*

** Corresponding Author*

In addition to the 14,422 deletions genotyped on the autosome and X, an additional 36 sites were genotyped on chromosome Y. The site selection criteria and genotyping methods employed were the same as for the autosome with the following differences:

a) (Y-specific read depth normalization) Read depth normalization was performed relative to the region Y:1-28780000 rather than relative to the whole reference genome. This was done to address problem with some samples that appear to be deficient in Y relative to the rest of the genome (perhaps due to mosaicism within the cell lines) and to avoid the repetitive region Yq12.
b) (genotype in males only) Samples labeled as being female were not used in the genotyping.
c) (remove highly repetitive sites) Sites < 1% uniquely alignable sequence were not included. Uniquely alignable sequence is defined as the number of base positions where a 36 bp window centered over the base position is unique within the reference genome.

The structural variation genotypes for chromosome Y were not included in the integrated call set, but are available as a separate data file.

5.18 Integration of SNPs, short indels, and SVs into a single call set

Authors: Hyun Min Kang, and Goncalo Abecasis

Each consensus SNP, indel, and SV call set were merged into a single VCF by simply considering each variant as a point mutation. The variant is linearly ordered by the genomic coordinate primarily based on the leftmost position and secondly the variant type (in the order of SNP, indel, and SV). The ploidy within the interval of SV was ignored in this integration step.

SNP genotype likelihoods were calculated using the BAM-specific Binomial Mixture Model (BBMM) described in the Baylor Low coverage SNP calling section. Indel likelihoods were calculated as described in the Broad Low coverage Indel calling section. The calculation of deletion genotype likelihoods was performed using Genome STRiP as described in the earlier description of structural variant detection.

The merged genotype likelihoods were used as the input to run the BEAGLE software[55] with 50 iterations across all samples together. The resulting haplotypes were refined using a modified version of the THUNDER software[56] with 300 states chosen by longest matching haplotypes at each iteration in addition to 100 randomly chosen states. The approach of using BEAGLE-estimated haplotypes to initialize methods such as THUNDER and IMPUTE2 produces higher quality haplotypes than any of these methods alone (Bryan Howie, personal communication). Processing all samples together facilitates downstream analysis, and work in the related field of genotype imputation has shown that these methods perform well in multi-ethnic datasets[57,58].

For chromosome X calling in the non pseudo-autosomal regions, the male genotype likelihoods for heterozygous allele were assigned as the minimum possible value under the diploid model.

5.19 Post-hoc short indel filtering

Authors: Adrian Tan, Hyun Min Kang, Goo Jun, Adam Auton, Scott Devine, Heng Li, and Goncalo Abecasis

After integration, we identified that a subset of indels from the low coverage data having very high false positive rates. In particular, 10 samples showed an excessive number of singleton indels (~1,000 to 23,000) that are mostly 1 bp insertions. These are individuals NA12144, NA20752, NA18626, NA18627, NA19313, NA19436, NA19437, NA19439, NA19446, and NA19448. Upon further investigation, we found that the excessive 1 bp singleton insertions are due to technical artifacts introduced in a specific cycle of the sequencing step in a particular run. We removed 162,928 1 bp singleton insertions specific only to these 10 outlier samples.

In addition, we found a much higher fraction of frameshift indels in low coverage specific indels compared to the indels shared between low coverage and exome data, suggesting that low coverage specific coding indels may have enriched false positive rates. We removed an additional 3,014 protein-coding frameshift indels exclusive to low coverage samples to increase the specificity of the protein-coding indels.

Preliminary evaluations of indel call sets demonstrated high apparent false positive rate after the above steps (~30% estimated false positive rates compared to Affymetrix exome array genotypes), and rare indels demonstrated higher discordance with independent datasets. To extract high quality indels, we restricted the minimum allele frequency (before integration) to 0.5%, and additionally applied SVM approach to further filter out potential false positive indels guided by the indel genotypes from the Affymetrix Axiom exome array genotyping chip were provided. The SVM was trained using multiple features including (a) allele balance (b) inbreeding coefficient (c) flanking sequence complexity (d) homopolymer runs (e) strand bias (f) cycle bias (g) mapping quality (h) number of supporting non-ref reads, and (i) distance to nearby indels. After filtering, estimated false positive rates from Affymetrix array were 5.4% (with the caveat of underestimate due to potential model overfitting).

6 Variant calling on chromosome Y

*Authors: Yali Xue, Yuan Chen, Shane McCarthy, Qasim Ayub, Luke Jostins, Richard Durbin, and Chris Tyler-Smith**

** Corresponding Author*

Calls were made on the 525 male samples in the Phase I release, plus an additional sample belonging to Haplogoup A (NA21313). Calls were made using samtools[21] and bcftools 0.1.17 (r973:277). Samtools was used to generate all-site, all-sample BCFs (samtools mpileup -DS -C50 -m2 -F0.0005 -d 10000 -P ILLUMINA). Sites were identified by calling the four continental groups, AMR, AFR, ASN and EUR, separately with bcftools (bcftools view -p 0.99 -bvcgs), then combining and recalling at the sites discovered in these four groups. The -s option in bcftools was used to identify the samples as haploid for calling. Calls were then merged and filtered (StrandBias 1e-5; EndDistBias 1e-7; MaxDP 10000; MinDP 2; Qual 3; SnpCluster 5,10; MinAltBases 2; MinMQ 10; SnpGap 3).

A revised version of Yfitter[6] was used to assign a haplogroup to each sample based on the variable sites called.

Site filtering and site QC matrix

Subsequent analyses concentrated on unique regions of the Y chromosome: 2,649,807-2,917,723; 6,616,752-7,472,224; 13,870,438-16,095,786; 16,170,614-17,986,473; 18,017,095-18,271,273; 18,537,846-19,567,356; 21,032,221-22,216,158; 22,513,120-23,497,661; 28,457,993-28,806,758.

We used half of the known variable sites reported in the literature[59] as gold standard sites, to produce a distribution of the read depth and genotype quality for these sites. We then applied a cutoff of read depth of 4-18x and genotype quality of 16-99 for at least one alternative allele call in that site to filter the sites in the raw vcf file. Then we used the remaining half of the literature sites to estimate the false negative rate of the filtered sites. 14 male individuals overlapped with Complete Genomics sequencing data[60]. We also used the Complete Genomics calls to estimate the false negative and false positive rates of the sites from the discordances between the callsets.

Genotype filtering and genotype QC matrix

Using the haplogroup assigned to each sample, we carried out the filtering steps described below.

1. For sites that appeared multiple times in one haplogroup but as singletons in other haplogroups inconsistent with the phylogeny, we treated the singletons as missing data (bad calls).
2. We treated sites that were found in two or more haplogroups with two or more calls each, but were inconsistent with the phylogeny, as recurrent sites and filtered them out when the phylogenetic tree was constructed.
3. We used the overlap with Complete Genomics data to set read depth and genotype quality cut off filters for bad calls, and treated these as missing data.
4. Two individuals were outliers in the tree, NA12413 with the highest proportion of missing data, and NA18603 which for unclear reasons was placed far from the position expected from its HapMap3 genotype. These individuals were both filtered out.

286 individuals overlapped with HapMap3 and therefore have Y-SNP genotype data available. Genotype data were obtained with the Affymetrix Human SNP array 6.0 (interrogating 1,852,600 genomic sites) and the Illumina Human 1M single beadchip (1,199,187 genomic sites)[61], and were used to assess the genotype accuracy.

Results:

In total, 18,699 unique region Y-SNPs were called in the raw VCF file. Only seven sites were filtered out using the quality cutoff determined using the literature sites, suggesting that the site calling quality is good. The site false negative rate was 17.2% (25/145) based on the literature sites, and 17.3% (393/2661) based on the Complete Genomics calls. The proportion of sites called in the 1000 Genomes analysis but not by Complete Genomics in the overlapping samples (maximum false positive rate) was 1.72% (142/8,269).

Among the filtered 18,692 Y-SNP sites, an ancestral state could be assigned for 16,679 (the allele matching the Ensembl release 66 chimpanzee Y chromosome sequence). We identified 720 recurrent sites, which were not included in the phylogenetic tree, and assigned 3,263 genotypes (0.03%) as missing data. The genotype accuracy was estimated at 97.4% (20,687/21,235) by comparison with the HapMap3 Y genotype calls.

7 Variant calling for mtDNA

*Authors: Hanjun Jin, Ki Cheol Kim, Wook Kim, Petr Danecek, Yuan Chen, Qasim Ayub, Yali Xue, and Chris Tyler-Smith**

** Corresponding Author*

For calling variants in the mitochondrion, a custom 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). Duplicate reads were removed by MarkDuplicates, implemented in Picard v1.36[8]. For subsequent analyses, we used the SAMtools package[21] to generate pileup files. Consensus sequences were then generated based on the pileup files by using SAMtools mpileup, bcftools view and vcfutils.pl vcf2fq commands from the SAMtools package. Indels were checked manually later. For all samples in this analysis, positions where the non-reference allele (compared with the revised Cambridge Reference Sequence (rCRS[62]) was covered by less than two reads were considered as 'N' (missing site and ambiguous site).

We excluded samples that had more than 1% Ns, or Ns in positions critical for haplogroup assignment, leaving 1074 samples. Mean coverage for each individual ranged from 53x to 7,555x and mean coverage for mtDNA sites ranged from 201x to 2,205x.

Heteroplasmy was called conservatively, with a mean MAF of 33%. The pattern of heteroplasmy is mostly even along the mtDNA molecule, with peaks within the control region as noted before[6].

8 Variant annotation

8.1 Functional annotation

Authors: Suganthi Balasubramanian, Ekta Khurana, Lukas Habegger, Arif Harmanci, Cristina Sisu, and Mark Gerstein

Coding annotations are based on the GENCODE7 gene annotation model[63]. This file was parsed to include all transcripts with a CCDS tag, and all transcripts whose transcript_type was labeled as "protein_coding" or "polymorphic pseudogene". In the latter set, transcripts labeled 'mRNA_start_NF' or 'mRNA_end_NF' were not included. Transcripts tagged as candidates for nonsense-mediated decay were also not included. The annotations were obtained using Variant Annotation Tool[64].

Non-coding categories used to annotate the variants include ncRNAs, UTRs, transcription factor (TF) peaks, TF motifs, enhancers and pseudogenes. ncRNAs are further divided into miRNA, snRNA, snoRNA, rRNA, lincRNA and miscellaneous RNA. ncRNAs, UTRs and pseudogenes are obtained from Gencode v7[65]. TF peaks, motifs and enhancers are obtained from Encode Integrative paper release[66,67]. A conservative set of enhancer elements is used which consists of intersection of those obtained using combined ChromHMM/Segway segmentation[66] with distal regulatory modules obtained by discriminative training[68].

8.2 Annotation of ancestral allele

Author: Laura Clarke

The SNP ancestral alleles were derived from the Ensembl 59 comparative 32 species alignment[69]. The VCF files were annotated using the VCFtools 'fill-aa' script[37], with the ancestral allele recorded using the 'AA' INFO tag. The ancestral allele FASTA files used for this annotation are available for download[70].

9 Validation and data quality

In order to assess the quality of the Phase 1 SNP calls, a series of validation experiments were performed for both the low coverage and exome call sets. Multiple independent technologies including PCR-Roche 454 and PacBio RS sequencers and Sequenom MassARRAY were used so as to ensure that the results were not skewed by error modes from just a single platform.

9.1 Low Coverage SNP validations

Authors: Danny Challis, Jin Yu, Fuli Yu, and Eric Banks

All three technologies were applied to a selection of 300 SNP loci from Chromosome 20 that were potentially polymorphic in at least one of eight samples (HG00321, HG00577, HG01101, NA20800, NA19313, NA20296, NA19740, NA18861). These samples were randomly chosen from the collection of samples available at different production centers. A locus was eligible for validation only if it had evidence of the alternate allele in at least one sequencing read among the eight samples. This however does not necessarily mean that the site was called polymorphic in any of the eight samples. The SNP loci were randomly selected from eligible sites from the Chr20 of the initial Phase 1 SNP call set.

9.1.1 Sequenom and Pacific Biosciences validation

We used the Mass Spectrometry genotyping technology and utilized AssayDesigner v.3.1 software to design PCR and extension primers for low multiplex SNP assays. Two sets of validation designs were created, one using 100 base pair PCR amplicons and another set of complementary 600 base pair amplicons. SNPs were amplified in multiplex PCR reactions consisting of a maximum of 12 loci each. The volume of the PCR reaction was used in both the Sequenom MassArray protocol and pooled together for the Pacific Biosciences sequencing aspect of the validation.

Sequenom: Following amplification, the Single Base Extension (SBE) reaction was performed on Shrimp Alkaline Phosphatase treated PCR product using *iPLEX* enzyme and mass-modified terminators. A small volume (approximately 7 nl) of reaction was then loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl of matrix (3-hydroxypicolinic acid). SpectroCHIPs were analyzed in automated mode by a MassArray MALDI-TOF Compact system with a solid phase laser mass spectrometer. The resulting spectra were called by the real-time SpectroCaller algorithm and analyzed by SpectroTyper v.4.0 software that combines a base caller with the clustering algorithm. 246 sites had usable Sequenom genotyping data after excluding those assays that failed the design phase as well as those with call rates of less than 75%. The Sequenom validation was applied to 383 samples (including the 8 validation target samples).

Pacific Biosciences: The RS sequencer output was processed using the Broad Institute-GATK PacBio Processing Pipeline (manuscript in review[71]). 30 sites could not be genotyped because the sequencing read coverage was less than 20x over all eight samples; the average coverage for the remaining 270 sites was over 500x. Sites were called and genotyped using the GATK Unified Genotyper[20]. Pacific Biosciences sequencing was performed just for the 8 validation target samples.

9.1.2 Roche 454 validations

The PCR-Roche 454 validations that were carried out at the Baylor College of Medicine-Human Genome Sequencing Center (BCM-HGSC) included two experiments. In the first experiment each SNP locus was validated in a single sample randomly selected from the subset of the eight samples with direct evidence of the SNP. This strategy made it possible to validate not only the SNP but also the genotype in the validation sample. However, this strategy is susceptible to producing false-negative results if a mismatch is a true SNP in one of the eight samples (that is, not selected for validation), but a sequencing error in another sample (which happened to be selected for validations). This issue is addressed in the second experiment, where the DNA for all eight samples was pooled and validated across all 300 SNP loci. While this does not allow determination of genotype or even which sample harbors the SNP, it ensures the SNP locus will be validated if it is really polymorphic among the eight samples.

Primer design for the PCR-Roche 454 validation experiments were performed using the Primer3 based BCM Primer Pipeline. For SNPs that Primer3 failed to design suitable primers, an attempt to manually design the primers was made.

Using this approach, primers were successfully designed for 273 of the 300 SNPs. The amplicons had an average length of 377 base pairs. The amplicons were PCR amplified and the resulting PCR reactions were normalized. The amplified DNA were then pooled and sequenced on the Roche 454 sequencing platform, generating 255k reads with an average length of 256 base pairs in the first experiment, and 260k reads with an average length of 258 base pairs in the second experiment. After removing sites that failed PCR, the average read depth coverage for each validation site was approximately 670.

The SNP were then genotyped using the Atlas-SNP pipeline[72]. These reads were mapped to the human reference genome (Build 37) using BLAT[73], and then aligned to the amplicon sequence using CrossMatch[51]. If the total read depth was less than 5, the site was considered a PCR failure and no call was attempted. If it was less than 50, the result was flagged for manual review. If the SNP was found with a variant read ratio above a defined cutoff (10% for single sample validation, 3% for pooled samples validation) it was considered confirmed, otherwise the SNP was considered a false positive. Of the 273 SNPs for which primers were successfully designed, 260 produced good results (that is, 95%) and 13 were PCR failures in the first experiment. In the second experiment, 264 produced good results (97%) and 9 were PCR failures.

9.1.3 Consolidation of validation genotypes

Once all three validation experiments were completed, the results were combined to minimize any ambiguity. Any locus that failed to produce reliable results in one of three experiments was usually reliable in another experiment. All but 6 of the 300 loci had reliable results in at least one of the experiments (Table S4).

Although the validation experiment was designed using a preliminary chr20 Phase 1 SNP call set, the results have been applied to the whole genome call set. Due to adaptations in parameters and filtering used in the SNP calling process between whole genome and chr20 data, 3 of the 300 validated SNPs are not included in the final call set. In addition, there are 10 SNPs, which while they had direct evidence in at least one of the eight samples, were not actually called in any of the eight samples in the final call set.

When combining the validation results, 15 of the 300 loci had contradictory results where one experiment identified the locus as a true SNP and another identified it as an error. 8 of these contradictions were caused by the locus being confirmed in one sample in one experiment, and not confirmed in a different sample in a different experiment. These 8 SNP loci were accepted as true SNPs. The other 7 contradictions are due to apparent errors in the validation experiments. These 7 loci were resolved by simple voting, with each validation experiment giving a single vote on the correct classification. To give a more accurate estimate of the final call set's quality the 10 SNPs not called in a validated sample and the 3 sites not in the final call set have been excluded from the results.

9.1.4 Results

Of the 287 remaining SNP loci, 276 of them were confirmed as true SNPs, giving an estimated FDR of 1.8% (Table S4). The validation results were also analyzed by minor allele frequency (MAF). Of the common (MAF>0.05) and low frequency (MAF>0.01) SNPs validated, all 83 were confirmed as being true SNPs, indicating that the higher MAF SNP calls are extremely reliable. For singleton and rare (MAF<0.01, excluding singletons) SNPs, the estimated FDR is higher, but is still below 5%.

The full validation results are available for download[74].

9.2 Exome validation

Authors: Danny Challis, Jin Yu, Fuli Yu, and Eric Banks

9.2.1 Exome SNP validation

Three series of exome SNP validation experiments were carried out at BCM-HGSC on both the consensus and unique SNPs calls of the three different centers using single sample PCR-Roche 454 validation as described above in the low coverage SNP validation. A total of 417 SNPs on chromosome 20 were selected from these validation experiments and submitted for primer design. 412 of those were designed successfully. The amplicons had an average length of 302 base pairs. The amplified DNA were then divided in five pools and sequenced on the Roche 454 sequencing platform, generating a total of 1419k reads with an average length of 287 base pairs. After excluding amplicons that failed PCR, the average read depth of coverage for each amplicon is 1625. After mapping the reads to the human genome and applying the same quality control steps and cutoff as in low coverage single sample PCR-Roche 454 validation, a total of 354 sites were genotyped confidently.

The detailed exome SNP validation are available for download[74].

9.2.2 Exome consensus SNPs stratified by allele frequency

100 singleton SNPs, 50 of allele frequency (AF) <1%, and 50 of allele frequency (AF) >=1% were randomly selected in this validation experiment to represent the SNP allele frequency distribution of the Phase1 integrated genotypes call set. At least one and at most 5 samples were chosen for each site to prepare the sequencing library pools and a total of 188 sites were genotyped confidently. For this set of SNPs, the overall FDR is estimated to be 1.6% (Table S5). All SNPs in AF>1% are validated as true positives and the FDR of singleton SNPs is 1.1%. The SNPs in <1% bins has a higher FDR of 4.1%. This may be explained by the fact that only a small subset of samples for each SNP were selected in this validation and imputation errors are more likely to be concentrated in this bin.

9.2.3 Novel exome consensus SNPs

Since a large portion of exome SNPs are not found in the low coverage SNP calls or dbSNP135, 100 of these consensus SNPs were randomly selected for validation. For each site, at least one and at most 2 samples were chosen to prepare the sequencing library pools and a total of 86 sites were genotyped confidently. The FDR of this set of SNPs is 2.3% (Table S5).

9.2.4 Center-unique exome SNPs

To assess the SNPs called exclusively by different centers and not selected in the exome consensus, we randomly selected at most 20 unique SNPs from both the Illumina and SOLiD platforms of the three centers for validation. For each site, at least one and at most 2 samples were chosen to prepare the sequencing library pools and a total of 85 sites were genotyped confidently. The FDR of each set are shown in Table S5.

9.3 Loss of Function (LoF) SNP validation

Authors: Jin Yu, Mike Jin, and Fuli Yu

To assess the quality of LoF SNPs in the Phase 1 integration release, we first applied several filters to remove possible annotation artifacts, excluded the sites that were included in previous experiments and then selected the remaining ones for single sample PCR-Roche 454 sequencing experiment.

9.3.1 LoF SNP selection

To produce a set of high confidence LoF SNPs, a series of fairly stringent filters were applied to all SNP sites predicted to create a nonsense codon or to disrupt a splice site in the Phase1 integrated release[75]. These filters included: 1) remove all SNPs with estimated AC=0; 2) remove splice SNPs in non-canonical splice sites or in introns shorter than 10bp; 3) remove SNPs where the LoF allele is the same as the inferred ancestral allele; 4) remove SNPs that effect only a subset of known transcripts and 5) remove nonsense SNPs found in the last 5% of the coding region of the longest affected transcript. After applying above filters, a total of 3,697 LoF SNPs of the whole genome were selected, which we identify as a high confidence LoF SNP set. This set consists of 2,535 nonsense and 1,162 splice-disrupting SNPs. We further excluded the sites that had been selected for validation in the 1000 Genomes Pilot LoF validation experiments, sites in exome chip design, and those in dbSNP release 129. The 2,003 remaining LoF SNPs consisted of 1296 nonsense and 707 splice-disrupting sites.

9.3.2 PCR-Roche 454 validations

We aimed to validate as many of the 2,003 LoF SNPs as possible using single sample PCR-Roche 454 sequencing validation at Baylor College of Medicine – Human Genome Sequencing Center (BCM-HGSC). SNPs were validated where samples were available in the BCM-HGSC DNA inventory. For singleton, doubleton and tripleton SNPs, we picked all available samples. For SNPs with higher allele frequencies, we randomly selected one sample from the DNA inventory at the BCM-HGSC. In total, 1,481 sites in 642 samples (1183 singletons, 150 doubletons, 35 tripletons, and 113 with Allele Count > 3) were submitted for primer design.

Primer design for the PCR-Roche 454 validation experiment was performed using the Primer3 based BCM Primer Pipeline. For SNPs that Primer3 failed to design suitable primers, an attempt to manually design the primers was made. Using this approach, primers were successfully designed for 1,405 of the 1,481 SNPs. Amplicons had an average length of 317 base pairs. Three pools were prepared and sequenced on the Roche 454 sequencing platform, generating a total of 2,005,944 reads with an average length of 261 base pairs.

Genotypes were called using the BCM-HGSC Atlas-SNP pipeline[72]. Of the 1,405 SNPs for which primers were successfully designed, 53 were PCR failures and 1,352 produced good results with an average read coverage depth of 1088 reads/site.

9.3.3 Results

The overall FDR of these 1,352 SNPs is 5.2%. Stratified by allele frequency, most of the LoF SNPs are singletons and doubletons, which also have the lowest FDRs of 3.1% and 4.8% respectively. With increasing allele frequency, the number of LoF SNPs decreases significantly, while both the FDR and 'No call' rate increase. There are only 8 LoF SNPs with allele frequency > 5%, although the FDR is as high as 80.0% (Table S8). This is in contrast to the results in exome validation experiments, in which the SNPs with allele frequency > 1% have an estimated 0% FDR (Table S5).

We propose several reasons for this phenomenon: 1) The LoF variants tend to be rare due to negative selection, the high frequency LoF SNPs are more likely to be artifacts. The high no call rate of the SNPs in this category also suggests many of them are in regions with low mappability; 2) Common LoF variants are more likely to be included in exome chip design and dbSNP129. These sites were excluded in this validation experiment, and hence the remaining sites are more likely to contain false positives; 3) As SNPs with allele frequency >=1% were only validated in one sample, conflation with genotyping error would lead to an increased false positive rate estimate.

The validation results are available for download[74].

9.4 Short indel Validation

Authors: Guillermo del Angel, Mauricio Carneiro, Eric Banks, Ryan Poplin, Namrata Gupta, Scott Donovan, Andrew Crenshaw, Liuda Ziaugra, Michelle Cipicchio, Melissa Parkin, Xinyue Liu, Ankit Maroo, Luke J. Tallon, Jeremy Gollub, Jeanette P. Schmidt, Christopher J. Davies, Brant A. Wong, Teresa Webster, Adrian Tan, Goo Jun, Hyun Min Kang, Mark DePristo, and Scott E. Devine

To assess the quality of the genome-wide Phase I indel call set, 93 indel sites were subjected to validation on three independent platforms: a) Sequenom mass-spectrometry-based genotyping, b) Pacific Biosciences (PacBio) targeted re-sequencing, and c) Roche 454 targeted re-sequencing. These three platforms together provided a more comprehensive view of indel validation than data collected from any of the single platforms alone. The 93 sites were examined in eight samples (HG00321, HG00577, HG01101, NA18861, NA19313, NA19740, NA20296, NA20800). Sites were chosen somewhat randomly with the following criteria: The site had to be polymorphic in at least one of the eight validation samples. To prevent a bias towards common alleles, sites also were chosen in a manner that retained the same allele frequency spectrum as the original input set. In a separate set of experiments, indels were examined on a custom Affymetrix Axiom array. A final high quality indel call set (5.4% FDR) was generated using a series of downstream filtering steps.

9.4.1 Sequenom and PacBio validation

Sequenom and PacBio indel validation was carried out at the Broad Institute. AssayDesigner v.3.1 software was used to design PCR and extension primers for low multiplex Sequenom indel assays. Two sets of validation designs were created, one set using 100 base pair PCR amplicons and the other set using 600 base pair amplicons. Each set of indel amplicons was amplified in multiplex PCR reactions consisting of a maximum of 12 loci. The 600 bp amplicon PCR reaction was used not only in the Sequenom MassArray protocol but also was used for the PacBio sequencing aspect of the validation. Sequenom assays were then analyzed by SpectroTyper v.4.0 software. The 600 bp amplicons also were used for PacBio re-sequencing, using target coverage of 120x at each site.

9.4.2 454 validation

454 indel validation was carried out at the Institute for Genome Sciences, University of Maryland School of Medicine. An independent set of primers was designed using Primer 3 software[76] with target melting temperatures of 63°C and a target PCR amplicon size of 400 to 600 bp. The 93 sites were amplified in the eight samples plus a negative control that lacked DNA. PCR products were evaluated on 2% agarose gels to ensure accurate amplicon sizes and to confirm that the negative controls lacked products. A total of 77/93 (80.5%) of the attempted PCR assays were successful and yielded robust products of the expected sizes. The PCR products were labeled with sample-specific bar codes, pooled, and sequenced with 454 Titanium chemistry. Sequencing yielded an average depth of 194 reads per indel allele and these were mapped to a library of all possible alleles to determine the genotype of each indel in the eight samples. Alignments were manually inspected to evaluate mapping and allele calling.

9.4.3 Results of PCR-based validation

Data from the Sequenom, PacBio, and 454 validation experiments were integrated and used to calculate a combined FDR of 35% for the exercise (Supplemental Table S6). Despite the fact that three independent platforms were used for validating indel calls, 17 indels remained uncalled by any platform.

These sites generally fell within simple repetitive regions that were difficult to analyze and/or did not amplify well in the PCR assays. On the basis of these data, we noted that indels generally fell into two broad categories in humans: 1) variants in "well-behaved" regions of the genome (regions containing complex sequences) with high quality mapping, sequencing, and alignment data underlying the variant calls, and 2) variants in more challenging regions of the genome (often within simple repeats) where lower quality mapping, sequencing, and/or alignment data may contribute to incorrect variant calls.

9.4.4 Axiom Exome genotyping array

The Affymetrix® Axiom® Exome Genotyping Array includes glass-bound, 30-mer oligo probes for 318,983 genetic variants, including SNPs from the Exome Chip Design Consortium[77], indels discovered in early versions of 1000 Genomes Phase I exome sequencing and low pass sequencing, as well as additional non synonymous coding SNPs from the Axiom™ Genomic Database. A total of 260,889 of these variants (including ~17,524 indels in exons and CDS regions) are expected to cause nonsynonymous changes in protein sequences. An additional 13,328 variants are predicted to cause synonymous changes. An additional 17,610 indels were identified in the 50 – 100 bp flanking regions of exons that were captured for exome sequencing. Additional variants were selected for significance in previous GWA studies, usefulness as ancestry informative markers or for calculation of identity by descent, and for other purposes described on the above website.

A total of 1,249 individuals from 14 populations were genotyped on the array using the Axiom 2.0 assay, which employs a ligation-based approach to interrogate whole genome-amplified DNA. Genomic DNA from the HapMap and 1000 Genomes collections was purchased from Coriell (population/DNA plate: CEU/T01, CHB+JPT/T02, CHS/MPG00002, CLM/MPG00005, FIN/MPG00001, GBR/MPG00003, IBS/MPG00010, JPT/MPG00009, LWK/MPG00008, MXL/MPG00006, PEL/MPG00011, PUR/MPG00004, TSI/MPG00007, YRI/T03). Genotyping analysis was performed with Affymetrix Power Tools software version 1.14.4. Genotype calls are available on the Affymetrix and 1000 Genomes websites.
In order to evaluate the FDR of the indels on the array that were derived solely from 1000 Genomes Phase I data sets, we limited our analysis to the genotyping data that was generated from 865 samples (exclusively Phase I data set-derived samples). Our analyses revealed an FDR of 33% - similar to the 35% FDR obtained with the PCR-based validations.

9.4.5 Short indel filtering

The indel validation experiment revealed a subset of indels contributing the majority of false-positive calls. In order to produce a final high quality indel call set, the calls were filtered as described in Section 5.19. After filtering, the estimated FDR for the final integrated low coverage indel call set was 5.4% (with the caveat of underestimation due to potential model overfitting).

9.5 SV validation

Three separate methods (SNP array intensity, PCR and aCGH) were used to validate the deletions in the integrated call set, although only the SNP array intensity evaluation, described below, was available when initial site selection for genotyping was made.

9.5.1 SV validation using Omni 2.5 SNP genotyping arrays

Authors: Robert E. Handsaker, and Steven A. McCarroll*

** Corresponding Author*

We used intensity data from the Omni 2.5 SNP genotyping arrays, which were run on every project sample, to perform validation of the putative deletion sites called by different calling methods. Although arrays (and individual array probes) can vary in their quantitative response to variation in copy number, the validation method employed is based on the simple idea that samples with lower copy number should, on average, exhibit lower signal intensities at a given probe than samples with higher copy number. Each deletion (or duplication) site is specified by chromosome and start and end coordinate and a set of samples thought to carry the deletion (duplication) at that site. The normalized probe intensities for each SNP were summed to create a single intensity value at each SNP position. We performed a non-parametric test that computes a Wilcoxon rank-sum *P*-value across all array SNP positions that underlie a given deletion or duplication site. The samples are first ranked separately at each position in intensity space and then the ranks across all positions underlying the putative deletion or duplication are used to calculate the Wilcoxon *P*-value that the samples carrying the deletion (duplication) have ranks below (above) the remaining samples. The FDR for a set of deletion or duplication calls was estimated as two times the fraction of putative deletion or duplication sites for which we measured a Wilcoxon rank-sum *P*-value $P > 0.5$.

In addition to using the SNP genotyping arrays to estimate FDR for sets of deletion calls, we also used a threshold of $P < 0.01$ for selection of individual sites for genotyping.

9.5.2 SV validation using PCR

Authors: Adrian Stütz, Sarah Lindsay, Matthew Hurles, and Jan Korbel

PCR validation experiments for deletions were designed using a spanning primer strategy where both primers hybridize to regions flanking the predicted SV. PCRs resulted in either a band size corresponding to the reference allele, or a shorter amplicon corresponding to the reference allele band size reduced by the inferred SV size. Each PCR was carried out along with two controls: NA12892 genomic DNA (control 1) and a pool of five DNAs, corresponding to four human samples (HG00407 + HG00689 + NA18507 + NA19314 (Coriell) and a chimpanzee sample EB176 (JC) (HPA Culture Collections) (control 2).

9.5.2.1 Design of PCR validation experiments

Random locus selection: To enable the calculation of FDRs for independent SV callsets, we randomly picked 96 loci from each deletion 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. Duplicate primers between different callsets were removed, yielding 91-96 loci tested per callset.

Primer design: We used an iterative PCR primer design pipeline to ensure the specific placement of primers into unique regions within 150 bp windows flanking the inferred SV breakpoint region (extended by the confidence interval, if available). The primer3 algorithm[76] was used for primer placement, with the option to "exclude primers matching onto known repeats". In-silico PCR[78] was applied (default parameters) with these primers 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.

9.5.2.2 PCR experimental conditions

PCR primers were synthesized by Sigma. PCR was carried out with JumpStart REDAccuTaq LA DNA polymerase (Sigma-Aldrich) on a PTC-225 DNA Engine Tetrad Cycler (Bio-Rad) in 25 ul reaction volumes. A first PCR experiment used the following parameters with 10 ng genomic DNA as template. Initial denaturation at 96°C for 30 sec; then 28 cycles of 94°C 5 sec, 58°C 30 sec, 68°C 8 min; followed by an additional cycle of 68°C for 30 min. A second, independent PCR experiment used these modified conditions, with 20 ng genomic DNA as template. Initial denaturation at 96°C for 1min; 5 cycles of 94°C 15 sec, 64°C 30 sec, 68°C 8 min; 5 cycles of 94°C 15 sec, 62°C 30 sec, 68°C 8 min; followed by 22 cycles of 94°C 15 sec, 60°C 30 sec, 68°C 8 min and an additional cycle of 68°C for 10 min. All PCR products were run on 1% agarose gels for 2h at 150V for band visualization and compared against the DNA ladders Hyperladder I and IV (Biolane). Selected PCR reactions were repeated and capillary sequenced with PCR primers from both ends.

9.5.2.3 Analysis of PCR validation data

Amplicons of both the test and control DNAs were analyzed by comparison to molecular markers without prior knowledge of expected band sizes. We recorded instances where only one, or both, alleles were observed, and where amplicon patterns were identical between sample and controls. Two independent PCR experiments were carried out and the results were merged. PCR results were not considered in cases where replicates were contradictory.

9.5.3 SV validation using custom CGH microarrays

Authors: Marcin von Grotthuss, Xinghua Mindy Shi, and Ryan Mills*

** Corresponding Author*

We designed a custom Agilent 2x1M CGH Microarray platform to assess the specificity of the CNV discovery algorithms. We coalesced the deletion and duplication calls across a 25-sample subset from the phase 1 sample list and segmented overlapping calls into Distinct Regions of Overlap (DROs). Each DRO was allocated 1 to 7 CGH probes depending on the size of the region. Preference was given to Agilent Catalog probes which fell in each DRO, however in many cases one or more custom probes were used. Custom probes were designed as follows: random oligomers of 45-60bp were considered in each region and scored for an optimal melting temperature (Tm) of 76, absence of simple repeats and also long homopolymer stretches. Candidate probes that mapped to the reference genome at more than 10 positions were discarded, as well as those that fell below an arbitrary minimum score. The remaining best scoring probes for each region were then utilized up to the maximum considered.

Custom Agilent 2x1M array-CGH DNA microarrays were used to validate deletions in 25 selected samples. The high-resolution probe design allowed for the direct interrogation of probes falling into predicted SV candidate regions. All probe-level data from the array were normalized using the default settings of Agilent’s Feature Extraction 10.10 software. Probes with saturated or varied reference signal intensity across array-CGH experiments were masked out as potentially non-specific or noisy. Such probes were identified if the mean log2 reference signal intensity was >16 or if the standard deviation of these log2 values was >0.6. Additionally, we excluded probes with high reference signal intensity and low log2-ratios across all samples, since such probes were likely also saturated and non-specific. The cut-offs used for the identification of these probes were derived empirically and are as follow: mean log2 reference signal intensity >10, absolute mean log2-ratio <0.4, and standard deviation of the log2-ratios <0.25. Log2-ratios values of non-masked-out probes were normalized by GC-content and each chromosome's median-shift. Deletion calls containing only a single probe post-filtering were omitted and did not contribute to the FDR calculation. For each other deletion, we selected a subset of probes that were deemed the most suitable to validate a call. This step was necessary, as the boundaries of some calls may have been overestimated resulting in the incorrect usage of distal probes that would be inappropriate for a call validation. We estimated that optimal probes were those that (i) had variable log2- ratios across the samples, which would reflect copy number differences, and (ii) whose log2-ratios were correlated in probe pairwise comparisons, which would indicate a coherent signature of a copy number variation. Therefore, for each deletion call probes were hierarchically clustered with the goals of maximizing both 58 the range of mean log2-ratios between the samples as well as the average pairwise Pearson correlation of the log2-ratio values of probes clustered. The clustering score was defined as the product of the first factor multiplied by the second. We considered a cluster of probes to be representative for a call if it contained 50%

or more post-filtered probes and if the average pairwise correlation of the log2-ratio values was >0. If none of the clusters met these two requirements, the call was marked as "NA" and was not used for the FDR estimation. If two or more clusters of probes could be considered as representative, all such clusters were used independently in the next step and the final decision as to which was most optimal was determined at the end of the validation process. The validation was performed by genotyping each locus and comparing our results with the ones provided in discovery sets. The following rules were applied to determine copy number variations:

- If, regardless of a cluster of probes used, the maximum of absolute mean log2-ratios was <0.5, we assigned two copies to all samples, unless the next rule was true.
- If the maximum of absolute mean log2-ratios was within the range of [0.35, 0.5), and the site was supported only by 3 or less post-filtered probes, the deletion call was marked as "NA", since it was likely the signal was not significant enough due to limited number of probes.
- If the maximum of absolute means was ≥0.5, sigmoid transformations of the means were modeled by *k-means* clustering (with *k*=2) and the means, within each cluster, were subjected to an Anderson-Darling normality test (using empirically derived *alpha*=0.01). The limited number of samples (25) precluded the modeling of the probes at each locus as a mixture of Gaussian densities. The sigmoid transformation, also known as a hyperbolic tangent (tanh), was applied to reduce the influence of extreme means on modeling, while the Anderson-Darling test was used to avoid the modeling of nonnormal densities. Copy number states [1, or 2] were assigned respectively to the samples.
- If the range of mean log2-ratios was ≥1.0, then the rule above was used but with *k*=3, and [0, 1, or 2] copy number clusters assignments.
- If two or more alternative models were created, that were differed by the *k* value and/or set of probes used, we chose as the most likely the model with the lowest clustering score. The clustering score was defined as a root mean square deviation of tanh mean log2-ratios from the cluster centers.

FDRs were calculated per-sample and were defined as the number of false discoveries over the sum of false and correct ones. If a deletion call was genotyped by us as a homozygous or heterozygous variant, then the call was evaluated as a true discovery; otherwise we marked it as a false discovery.

10 Analysis

10.1 Quantifying the Phase 1 dataset

Author: Mark DePristo

For the data presented in Table 1, the Phase 1 integrated haplotypes were first partitioned into variants present in the autosomes or the X chromosome. Variants were categorized as SNPs if they represent single base length-preserving substitutions with respect to the reference allele, indels if they imply a change in the size of the genome sequence of 50 bp or less, and CNVs otherwise. Per-sample averages for each variant type were determined by considering a variant present in a sample if its corresponding genotype included at least one non-reference allele in the reference panel. SNPs and indels were identified as synonymous, non-synonymous, and nonsense or in-frame or frameshifting, respectively, according to their corresponding functional annotations (nonsense includes annotations prematureStop or removedStop, while frameshift includes only deletionFS or insertionFS versus deletionNFS and insertionNFS). SNP and indels were considered novel if their left-aligned start position did not overlap with an variant in dbSNP 135 (excluding sites uniquely identified in the preliminary 1000 Genomes Phase I release included in build 135). CNVs were considered novel if the CNVs has <50% reciprocal overlap in its start and end position on the genome with any CNVs classified as human, germline SVs in dbVAR (March, 2012).

10.2 Assessment of power of variant discovery and genotype accuracy

Author: Hyun Min Kang

The assess the power to detect variants in Phase 1, individual genotypes from Omni2.5M SNP array are compared to the Phase 1 integrated genotypes. The power is defined as [# variants with positive 1000 Genomes variant count]/[# variants with positive Omni2.5 variant count], grouped by the Omni2.5 variant count among 1,092 individuals. Whole genome power was evaluated across the genome. To avoid the underestimation of power in the exome data, the exome power was evaluated within consensus target using the genotypes concordant between Omni 2.5M SNP array and Affymetrix Exome Array genotypes. The resulting estimates are show in Figure 1a.

We estimated that 98.3% of SNPs for each individual are included in the integrated callset by comparing with OMNI2.5 genotypes. This estimate was obtained by first estimating the expected allele frequency spectrum (AFS) across whole genome from the AFS of synonymous variants within the consensus target region, after adjusting for the false negative rate evaluated by OMNI2.5 (as shown in Figure 1a). We then calculated the per-individual SNP discovery rate by comparing OMNI2.5 SNP genotypes with whole genome integrated genotypes, weighted by the ratio of expected to observed whole-genome AFS across all possible variant count.

To estimate genotype accuracy, the squared Pearson's correlation coefficient (r^2) between Omni2.5 array genotypes (observed variant count per individual) and the genotype dosages (expected variant count per individual) was calculated for each SNP across the 1,092 individuals. The average r^2 value (among 1,092 individuals) was calculated across the genome (WGS), and within the consensus target (Exome). The genotype dosages were calculated from the posterior probability of the MaCH/Thunder haplotyping software (with LD), or from posterior probability computed from genotype likelihood and estimated allele frequency (without LD). The resulting estimates are shown in Figure 1b.

10.3 Variant discovery by low coverage and exome sequencing

Author: Erik Garrison

To assess the relative contribution of the low coverage and exome sequencing towards variant discovery in the exome capture targets (Figure 14), we first extracted from the integrated callset the subset of sites contained within the exome consensus targets. The integrated calls retain information describing the experiment from which each variant was discovered, either exome, low coverage, or both. We used this source information to plot the relative fractions of the contribution of the exome and low-coverage sequencing projects to the integrated set by alternate allele count in the 1092 genomes.

10.4 Assessment of the accessible genome in Phase 1

Author: Goncalo Abecasis

Due to the nature of short-read sequencing, the sequencing depth varies along the length of the genome. As such, not all regions of the genome will have equal power for variant discovery. To assess provide an assessment of the regions of the genome that are accessible to the next-generation sequencing methods used in Phase 1, we created two genome masks.

Most project analysis did not use these hard masks for calling. Instead, the project used the VQSR algorithm (implemented in GATK) to distinguish variants likely to be true positives from others more likely to be false positives. However, the masks are useful for (a) comparing accessibility using current technologies to accessibility in the pilot project, and (b) population genetic analysis (such as estimates of mutation rate) that must focus on genomic regions with very low false positive and false negative rates.

Two sets of masks are available – a 'Pilot-style' mask and a 'Strict' mask. Each base in the genome is coded as follows:

- N - the base is an N in the reference genome GRCh37
- L - depth of coverage is much lower than average
- H - depth of coverage is much higher than average
- Z - too many reads with zero mapping quality overlap this position
- Q - the average mapping quality at the position is too low
- P - the base passed all filters
- 0 - an overlapping base was never observed in aligned reads

Regions marked as N, L, H, Z, or Q are less accessible to short reads. Although they can still be analyzed they are more prone to false positives.

The Pilot-style mask was produced using the same definition as used in the 1000 Genomes Project Pilot Paper[6]. This definition excludes the portion of the genome where depth of coverage (summed across all samples) was higher or lower than the average depth by a factor of 2-fold. It also excludes sites where >20% of overlapping reads had mapping quality of zero. The average total depth of coverage across Phase I samples is 5132. Thus, sites with a depth of coverage of <2566 or >10264 were excluded. Since approximately one half of project samples are males, depth of coverage is generally lower on the X chromosome. Coverage thresholds on the X were adjusted by a factor of 3/4.

Overall, this Pilot-style mask results in about 6.6% of bases marked as N, 1.4% marked L, 0.4% marked H and 3.9% marked Z. The remaining 87.8% of passed are marked passed (P) - and correspond to 94.0% of non-N bases.

As the name suggests, the Strict mask uses a more stringent definition. This definition uses a narrower band for coverage, requiring that total coverage should be with 50% of the average, that no more than 0.1% of reads have mapping quality of zero, and that the average mapping quality for the position should be 56 or greater. This definition is quite stringent and focuses on the most unique regions of the genome. In the regions that are marked as passed by this mask, only ~2% of sites called in an initial analysis are marked as likely false positives by VQSR. The average total depth of coverage across Phase I samples is 5132. Thus, sites with a depth of coverage of <2566 or >7698 were excluded.

Overall, this strict mask results in about 6.6% of bases marked N, 1.4% marked L, 0.9% marked H, 22.1% marked Z, and 1.6% marked Q. The remaining 67.5% of passed are marked passed (P) - corresponding to 72.2% of the non-N bases.

Each mask is summarized in both a FASTA-style file and a BED-style file, which are available for download[79].

10.5 Haplotype estimation from OMNI data

Authors: Olivier Delaneau and Jonathan Marchini

Haplotypes were estimated from the Illumina OMNI genotypes on 2,123 samples in the following way.

The genotypes were converted to PED/MAP format using VCFtools[37]. SNPs with the following entries in the FILTER column were excluded: amb, dup, id10, id20, id5, id50, refN. Family relationships were derived from the files in the sample pedigree file[80].

The resulting dataset contained 327 trios, 42 duos and 1,058 unrelated to give a total of 2,123 individuals at 2,177,885 SNPs. A detailed breakdown of the numbers of trios, duos and unrelated samples in each population is given in Table S3.

We found two trios with very high Mendel error rates on specific chromosomes. The CEU trio parent NA06984 has very low heterozygosity (0.00086 in the region approx. 70-80Mb of chr18) and the ASW trio child NA19918 has very low heterozygosity (0.00015 in region approx. 0-7Mb of chr17). This is likely due to uniparental disomies in the cell line DNAs of these samples. When phasing chr17 and chr18 we ignored the familial relationships between the samples in these two trios.

The program SHAPEIT[81] was used to phase this dataset one chromosome at a time. This program can handle trios, duos, and unrelateds at the same time and has been shown to provide highly accurate solutions compared to all the most widely used phasing programs. The resulting files were converted to VCF format, and are available for download[82].

10.6 Imputation using the Phase 1 data

Author: Bryan Howie

To evaluate the utility of the Phase 1 haplotypes as a genotype imputation resource, we performed cross-validations in external datasets genotyped with different technologies. We measured imputation accuracy at SNPs, short insertion/deletion polymorphisms (indels), and large deletions from the reference sequence (structural variants, or SVs).

10.6.1 SNP and indel evaluation with Complete Genomics data

We assessed accuracy at SNPs and indels with high-coverage, whole-genome sequence data made publicly available by Complete Genomics, Inc. (CGI). Of the 69 individuals in this resource, 20 are neither included in Phase 1 nor related to Phase 1 samples. These include 9 individuals of African ancestry (3 LWK, 4 MKK, 2 YRI), 3 individuals of admixed American ancestry (3 MXL), 4 individuals of European ancestry (3 CEU, 1 TSI), and 4 individuals of south Asian ancestry (4 GIH).

To emulate a typical imputation analysis in an association study, we masked the CGI genotypes at all sites not included on an Illumina 1M SNP array and then imputed the masked genotypes from the pseudo-array scaffold and the Phase 1 integrated variant haplotypes. Imputation was performed by supplying the 20 CGI-sequenced individuals and the full set of Phase 1 haplotypes to IMPUTE2[83], which chooses a custom reference panel for each study individual in each 5-Mb segment of the genome. We set the k_{hap} parameter of IMPUTE2 (the number of reference haplotypes to use when imputing each individual) to 1500 since pilot experiments showed that this value provided high accuracy in all populations. All other software settings followed the default values of IMPUTE v2.2.2.

This experiment produced estimated genotype probabilities for all SNPs and indels in the Phase 1 reference panel. We measured imputation accuracy by comparing these estimates with the CGI genotypes at shared variants. Following standard practice in the genotype imputation field, we defined accuracy as the squared correlation between imputed allele dosages, which take values in [0,2], and masked CGI genotype calls, which take values in [0,1,2]. Imputed variants were assigned to bins according to the allele frequency of each continental group in the Phase 1 callset; since there are no south Asians in Phase 1, we calculated the frequencies for this group as a weighted sum of the European (w = 0.67) and east Asian (w = 0.33) frequencies at each variant. For each variant type, allele frequency bin, and ancestry group, we computed the squared Pearson correlation (R^2) between the aggregate dosages and masked genotypes. This differs from the standard approach of computing variant-wise correlations and taking the average within each frequency bin. Aggregate statistics are more stable than variant-wise means in small sample sizes like the ones used here, and our experiments show that aggregate correlations for small samples are similar to variant-wise average correlations for large samples (data not shown).

We used the CGI results to create allele frequency vs. imputation accuracy curves for three variant classes in each ancestry group: genome-wide SNPs, exome-wide SNPs, and genome-wide indels (Figures 5a, S14). Exome SNPs were defined as those that fall within consensus target regions of the capture arrays used for Phase 1 exome sequencing. We excluded 83 exome SNPs from this analysis due to excess heterozygosity (inbreeding statistic less than -0.95 across all Phase 1 individuals), which is a hallmark of spurious variants caused by segmental duplications that are not present in the reference sequence.

Our initial results showed that the apparent imputation accuracy of indels was consistently lower than that of SNPs. Comparisons with array-based indel genotypes suggest that this effect is driven by the difficulty of calling indel genotypes from short sequence reads in both the Phase 1 and CGI datasets (data not shown). To generate a comparison set enriched for variants that were called well in both datasets, we restricted the indel results to sites not located within the Phase 1 sequence mask.

10.6.2 SV evaluation with Conrad *et al.* data

We evaluated the imputation accuracy of large deletions (SVs) by comparing against tiling array genotypes from a large study of copy number polymorphism[84]. We simulated a SNP array by using HapMap 3 genotypes at sites on the Illumina 1M platform. The HapMap 3 and Conrad *et al.* datasets share 74 YRI (AFR) and 76 CEU (EUR) individuals who are not included among or related to Phase 1 samples. For these 150 individuals, we imputed 1,956 SVs that were genotyped in both Phase 1 and the Conrad *et al*. study and had at least 80% reciprocal sequence overlap. The imputation procedure followed the same parameters outlined above. We measured imputation accuracy as the aggregate squared correlation between masked Conrad *et al.* genotypes and imputed dosages within each allele frequency bin. The results are plotted in Figure 5a and S14, and they show that SVs can be imputed with accuracy similar to that of SNPs.

While this experiment mimics the imputation of SVs from the Phase 1 reference panel into an external dataset (e.g., an association study cohort), we can also evaluate SVs that were imputed *within* the Phase 1 data set. When genotype likelihoods were constructed for variant integration, the likelihoods for SVs were calculated only for individuals with Illumina sequencing data, comprising 944 of the 1,092 samples in the Phase 1 set. For the remaining 148 individuals, the genotypes in the integrated call set were imputed from the genotype likelihoods for nearby SNPs and INDELs. For 24 of these individuals (12 AFR, 6 EUR, 6 E.ASN), we assessed the accuracy of the genotypes in the integrated call set by comparing against array-based genotypes from Conrad *et al*. We measured the accuracy of within-Phase 1 imputation using the same 1,956 SVs described above.

The results are shown in Figure S14c, which shows the aggregate squared correlation (R^2) between Conrad et al. genotypes and Phase 1 genotypes as a function of Phase 1 allele frequency. As with imputation into an external dataset, we find that accuracy is high for common SVs imputed within the Phase 1 samples: all populations have $R^2 > 0.8$ for SVs with frequency greater than 5% and $R^2 > 0.9$ for SVs with frequency greater than 10%. This high level of agreement suggests that the variant integration process was generally successful for large deletions.

10.6.3 Comparison of Phase 1 haplotypes with benchmark haplotypes

To further evaluate the utility of the Phase 1 haplotypes as a genotype imputation resource, we compared them against a set of high-quality benchmark haplotypes. The benchmark haplotypes were generated by genotyping all Phase 1 individuals on the Illumina Omni 2.5 M SNP array, together with a number of family members and unrelated individuals collected for later phases of the project; 1,856 samples were genotyped in total. These array genotypes were phased across entire chromosomes by SHAPEIT[81], which can handle a mixture of unrelateds, duos, and trios. Of the 1,092 Phase 1 individuals, 380 were phased as trio parents (95 AFR, 169 AMR, 100 ASN, 16 EUR), 35 as duo parents (34 AFR, 1 AMR), and 679 as unrelateds (117 AFR, 11 AMR, 186 ASN, 363 EUR). The entire set of 1,856 samples was phased together while using transmission information from all known family relationships.

The SNP array genotypes have been shown to be accurate, and the genotyped family members in this sample set should yield high-quality phasing, so the phased Omni haplotypes provide a useful benchmark for assessing the Phase 1 haplotypes. For an apples-to-apples comparison, we reduced both datasets to the same set of individuals (1,092 Phase 1 samples) and variants (52,114 chromosome 20 SNPs typed on the Omni 2.5 M array and called in Phase 1). We then used each of these haplotype sets as a reference panel to impute masked genotypes in the remaining unrelated individuals typed on the SNP array; these included 46 AFR, 103 AMR, 88 E.ASN, 108 EUR, and 76 S.ASN samples.

When imputing the non-Phase 1 individuals, we masked every 25th Omni SNP in sliding windows, such that the SNPs were effectively imputed from a 2.5 M array scaffold and every Omni SNP was imputed exactly once. We imputed each masked SNP from two reference panels: the Phase 1 haplotypes and the Omni benchmark haplotypes. The masked genotypes were imputed by IMPUTE2 (version 2.1.2) on default settings.

The imputation accuracy (mean R^2 between masked array genotypes and imputed allele dosages) is shown in Figure S14b as a function of non-reference allele frequency for each broad ancestral group. The solid and dotted lines show the accuracy obtained when imputing from the Phase 1 and benchmark reference panels, respectively.

As expected, the Omni benchmark haplotypes provide higher imputation accuracy across most populations and allele frequencies. Nonetheless, imputation from the Phase 1 haplotypes achieves competitive accuracy in all cases, which is striking because these haplotypes were inferred primarily from low-coverage sequence data and without the benefit of genotyped family members. These results suggest that the Phase 1 haplotypes are of high quality and can be viewed as a reliable reference panel for genotype imputation in association studies.

It is possible to take advantage of the family-informed OMNI haplotypes when calling genotypes and phasing haplotypes from sequence data[85]. Subsequent imputation from such haplotype reference panels can lead to a substantial boost in imputation accuracy at rare variants.

10.7 Analysis of private and cosmopolitan variants by frequency

Authors: Adam Auton and Gil McVean

To generate Figure 2b, we extracted the alternative allele counts for all variants in the Phase 1 release. For each variant, we also calculated the alternative allele count in each population or continental grouping (AFR: ASW, LWK, YRI; AMR: CLM, MCL, PUR; ASN: CHB, CHS, JPT; EUR: CEU, FIN, GBR, TSI). Using this data, we determined if the alternative allele was private to a specific population / continent, or shared across multiple populations or continents ('cosmopolitan'). Figure 2b shows the fraction of sites restricted to single populations/groups and the fraction shared across each group/population.

Because rare variants will (whatever the true degree of differentiation) typically be found in only one or a few populations, we also calculated a metric of allele sharing within and between populations that can detect the excess of differentiation relative to chance. Specifically, we compute, for each group of populations, the probability of sampling (without replacement) two chromosomes carrying the variant allele if the chromosomes are drawn from the same population (weighted by the number of sample-pairs within each population). This is divided by the probability of sampling (without replacement) two chromosomes carrying the variant allele if the chromosomes are drawn from the entire pool of populations. If the number of copies of an allele in population *i* is a_i, the number of haploid genomes in that population is n_i, the total number of copies of the allele across the group is a and the total number of haploid genomes is n, then the statistic is

$$F = \left(\frac{\sum_i a_i(a_i-1)}{\sum_i n_i(n_i-1)}\right) / \left(\frac{a(a-1)}{n(n-1)}\right).$$

For a given value of a, the statistic is averaged over all sites in which the allele count in the group is a. This statistic was used to calculate the excess sharing shown in Figure S6a. The statistic can also be computed between groups (by pooling all sampling within a group), to give the black line in Figure S6a. The statistic can also be computed for a single population:

$$F_i = \left(\frac{a_i(a_i-1)}{n_i(n_i-1)}\right) / \left(\frac{a(a-1)}{n(n-1)}\right).$$

Again, the statistic can be averaged over all sites where the allele count in the wider group is a. These plots are shown in Figure S6b.

10.8 Density of variants as a function of derived allele frequency

Authors: Adam Auton and Gil McVean

In Figure 2c, we plotted the density of the expected number of variants per kilobase carried by a genome drawn from each population (i.e., the integral of this function gives the average number of variants (per kb) carried by a haploid genome drawn from the population. This expectation was calculated for a given allele count, j, as $\hat{\theta}_j = 1000 \times j\eta_j/(G\alpha\rho)$, where G is the genome size (taken as 2.85Gb), α is the fraction accessible (0.94), ρ is the fraction of variants where ancestral allele status can be assigned with high confidence (0.86), and η_j is the number of variants with frequency j.

10.9 Analysis of highly differentiated sites

*Authors: Vincenza Colonna, Yali Xue, Yuan Chen, Qasim Ayub, and Chris Tyler-Smith**

** Corresponding Author*

Derived allele frequencies (DAF) were calculated for each population or continent using data from the final integrated call set. Continents were AFR: ASW, LWK, YRI; ASN: CHB, CHS, JPT; EUR: CEU, FIN, GBR, TSI. We excluded populations with small sample size (IBS) and extensive admixture (CLM, MXL, PUR). For each pair of populations or continents ΔDAF was calculated for each SNP as the absolute difference between DAFs in each population or continent.

Highly differentiated (HD) sites tend to cluster in the genome. However, most likely only one or a few SNPs in each cluster have functional consequences that have driven the observed extreme differentiation. Therefore, from the highest 1% of each ΔDAF distribution a subset of sites was chosen according to two criteria: each SNP is the most highly differentiated SNP in every 1000 SNPs from non-overlapping chromosome intervals, and it must have ΔDAF ≥ 0.7 or ≥ 0.25

for continental and population comparisons, respectively. The filtered subset should be enriched for sites with functional consequences, and depleted of sites that do not contribute directly to the phenotype.

Validation was performed using two approaches. First, Complete Genomics (CG) data from five overlapping populations (ASW, LWK, YRI, CEU, CHB) with 868 overlapping sites was used to evaluate consistency of ΔDAF values. Fifty CG individuals were used after excluding closely related ones. Thirty-nine of them overlap with Phase I samples. Due to the small CG sample size, populations were pooled into continents (AFR, 2n=18; ASN, 2n=8; EUR, 2n=24) and we limited this comparison to the continental level. Second, the most highly differentiated HD sites from continental and population comparisons (n=696 sites) compatible with Sequenom assay design was chosen and genotyped in 362 Phase I individuals using a Sequenom assay, and used to evaluate genotype concordance, expressed as the ratio between concordant and total calls.

We identified between 17 and 343 HD sites in population comparisons and between 190 and 348 in continental comparisons (Table S12). Validation using CG data showed that ΔDAF values obtained from the CG dataset were consistently correlated with those calculated from the Phase I call set (Pearson's product-moment correlation EUR-ASN, r^2=0.81, p-value < 2.2e-16; AFR-ASN, r^2=0.79, p-value < 2.2e-16; AFR-EUR, r^2=0.79, p-value < 2.2e-16). Validation using Sequenom showed an average per-locus genotype concordance rate of 95% after removing sites where the Sequenom assay failed (n=604 sites remaining). The highest discordance rate was found in homozygote alternative calls (8.2%; 3.7% for homozygous reference; 4.8% for heterozygous).

There were no fixed differences between any pair of continents or populations, a finding we interpret as a likely consequence of shared ancestry and recent genetic exchanges either at population and continental level. ΔDAF values between populations were generally higher in AFR (median range: AFR = 0.16-0.19; EUR = 0.10-0.17; ASN = 0.10-0.15), and CEU-GBR have the lowest number of HD sites. These findings largely reflect population sampling choices, which took population similarity into account to different degrees in different continents, rather than biological properties.

The highest ΔDAF value at each level is found at a site already known to be highly differentiated and under selection: 0.98 between ASN-EUR at rs1426654 in the SLC24A5 gene, causal for light skin in Europeans at the continental level; 0.63 at rs4988235 located in the *MCM6* gene, and the promoter of the Lactase (*LCT*) gene associated with lactose intolerance at the population level. Besides these two, a number of other ‘known’ sites were identified, among which were *DARC*, *EXOC6B*, *DOK5*, *SLC24A5*, *SLC45A2*, *EDAR*, and *TLR1*.

10.10 Rare allele sharing within and between populations

Authors: Adam Auton and Gil McVean

In Figure 3a, we investigated patterns of allele sharing for variants with very low frequencies. Specifically, we identified all variants with a minor allele count of exactly 2 across the entire Phase 1 sample, corresponding to a frequency estimate of 0.09%. We refer to these as f_2 variants. For each f_2 variant, we tabulated the populations in which the two copies of the variant were contained, allowing estimation of the relative proportion of rare allele sharing between populations. Figure 3a summarises the results in a graphical form.

10.11 Shared haplotype length as a function of allele frequency

Authors: Dionysia K. Xifara and Gil McVean

To investigate the extent of haplotype sharing between variants of differing frequencies (Figure 3b), we performed the following analysis. We isolated 196 regions of length 1Mb, randomly sampled along the genome. For every segregating site within these regions, excluding the first and last 100 kb, we considered up to 15 pairs of haplotypes from the same population, randomly selected among all haplotypes that carried that variant (excluding samples identified as cryptically-related; see Table S10). We determined the distance from the current position to the *k*th site at which the haplotypes differed in either the 5' and 3' direction, recording the total length of the shared chunk. Experiments showed that while increasing *k* from 1 to 2 roughly doubled the length of shared haplotype (as expected if the 'break' is due to a genotyping error), increasing *k* from 2 to 3 had a much smaller effect, suggesting that at k = 2, haplotype identity is lost either through true breaks in identity, phasing switch errors or clusters of incorrect genotypes and false SNPs. We therefore report haplotype length to the second different allele between haplotypes. For each allele count, we report the median shared chunk length over all the regions, weighted by the number of SNPs that had been considered. Figure 3b illustrates the reduction in shared chunk length as allele frequency increases, for each population. Genetic distances were estimated from the combined-population fine-scale genetic maps estimated from the HapMap2 data (The International HapMap Consortium 2007). The expected curve for genetic distance (Figure S7a inset) was obtained assuming a coalescent model with exponential population growth, starting 10,000 years ago and increasing the effective population size from 10,000 to 4 million (after Nelson *et al*. 2012).

10.12 Local ancestry inference

Authors: Eimear E. Kenny, Claire Churchhouse, Anjali Gupta Hinch, Amy Williams, Yael Baran, Simon Gravel, Brian Maples, Fouad Zakharia, Eran Halperin, Simon Myers, Jonathan Marchini, and Carlos D. Bustamante*

** Corresponding Author*

For African American (ASW, n=61), Mexican (MXL, n=66), Puerto Rican (PUR, n=55) and Colombian (CLM, n=60) individuals, we inferred the continent of origin for each base pair along the genome using a common panel of African (AFR), European (EUR), and Native American (NAT) individuals serving as proxies for the ancestral populations. The AFR reference panel (n=198) comprised individuals from Yoruba (YRI, n=101) and Luhya (LKK, n=97), the EUR reference panel (n=395) comprised individuals from Great Britain (GBR, n=99), Tuscany (TSI=100), Iberia (IBS=97) and the CEPH-panel (CEU, n=99), and a NAT reference panel (n=45) was from Mao *et al.*[86]. To produce a 'high accuracy' call set for each admixed population we considered calls that were a consensus across multiple local ancestry inference method; LAMP-LD[87], HAPMIX[88] (personal communication, S. Myers), RFMIX (personal communication Bustamante) and MULTIMIX[89]. The resulting consensus set of local ancestry tract calls for all TGP phase 1 admixed individuals was generated for the project and is available from the FTP site[90].

Local ancestry inference calls were made using the OMNI 2.5M genotype data and low pass sequencing SNP calls available at the interim release of June 2011. Each local ancestry inference method takes as input phased genotype data for both admixed individuals and the putative ancestral panels. The SHAPEIT algorithm with default parameters[91] was used to phase all haplotypes. Local ancestry inference in ASW individuals used the EUR and AFR panels and the OMNI 2.5M genotypes. A third reference panel of NAT typed on the Affymetrix 6.0 genotype chip were used for local ancestry inference of MXL, CLM and PUR individuals, and we therefore constructed a set of genotype calls at Affy6 sites for all MXL, CLM, PUR, EUR and AFR individuals using TGP data. First, genotypes at Affy6 sites were extracted from the low-pass sequencing SNP calls and then merged with the subset of OMNI 2.5M genotypes that were at Affy6 sites. Trio structures can vastly improve phasing, so the children of the MXL, PUR and CLM trios were included, albeit just using the subset of OMNI 2.5M genotypes at Affy6 sites (therefore, the children of each trio had non-negligible quantity of missing data). Genotypes used for calling MXL, CLM and PUR are available on the FTP site. Finally, local ancestry tracts were called by each method for both haplotypes of each chromosome per admixed individual using the following parameters (i) LAMP-LD[87] using S=25 and W=100, (ii) HAPMIX[88] with default parameters for ASW and, for MXL, PUR and CLM, an extended version of HAPMIX that uses a constraint solver to combine results from three two-way runs, (iii) RFMIX using window sizes of 0.2 cM and assuming 8 generations since admixture and (iv) MULTIMIX with default parameters.

In order to generate the set of consensus calls, the per-chromosome haplotypes were collapsed at each genotyped site (Affy6 or OMNI 2.5M) to give single diploid call per site (probabilistic calls were rounded to 0 or 1) in all admixed individuals across all four methods. Diploid calls were generated to avoid errors in haploid calls due to switch errors in phasing. Calls across four methods were compared at every site to generate the 'high accuracy' consensus calls; where 3 or 4 of the methods agreed, the majority call was set as the consensus call, if there was a tie, 3- or 4-way disagreement between methods, the site was set to 'unknown'. Sequential diploid calls along chromosomes were collapsed to give diploid call tracts of base pair length the distance between the first and last SNP in the tract. Base pairs between the SNP at the end of one tract and the first SNP at the beginning of the next tract were labeled ‘undetermined’. Regions at the start and end of each chromosome flanking the first and last SNP sites used in the calling are also listed as ‘undetermined’. Global proportions of EUR, AFR or NAT ancestry in ASW, MXL, PUR and CLM admixed individuals and the proportion called as ‘unknown’ were calculated by averaging over all diploid ancestry called tracts (excluding ‘undetermined’ tracts; Figure S9a).

Once local ancestry assignments had been performed, we obtained the proportion of novel SNPs, heterozygous sites, and nonsynonymous-to-synonymous ratio at non-reference sites as a function of the diploid ancestry (AFR/AFR, AFR/EUR, NAT/NAT etc.). The proportion of heterozygous sites in each ancestry category (Figure S9b) was calculated among sites with sufficient data and passed quality filters (specifically using the Pilot-style mask described in section 10.4)[79]. The proportion of novel SNPs in each population (Figure 3c) was calculated with respect to the list of SNPs in dbSNP version 135 but pruned for SNPs that were previously discovered and reported by the 1000 Genomes Project. Finally, the rate of nonsynonymous-to-synonymous SNPs in each ancestry category (Figure S9c) was calculated over all sites that were not homozygous for the reference allele.

10.13 Estimation of F_{ST}

To compare the level of genetic differentiation between populations, we calculated the commonly used statistic, F_{ST}. There are many possible estimators of F_{ST}, each of which can give differing results. We note that the estimated values of F_{ST} can be quite sensitive to choice of estimator and the inclusion of rare variation.

10.13.1 Weir and Cockerham, HapMap estimators

Author: Adam Auton

We estimated F_{ST} for each pairwise population comparison for both the Weir and Cockerham[92] and the HapMap[93] estimators. This was achieved by using VCFtools[37] for each autosome separately. For a given estimator, the estimates were very similar for each chromosome, and we therefore calculated a combined estimate by averaging the per-chromosome values. The results are shown in Table S11. Although the two estimators provide different estimates, the relative ordering of pairwise comparisons is largely consistent, with the exception of the IBS population, which has a small sample size in Phase 1. These estiamtes can be demonstrated by the removal variants with MAF < 5% in the combined sample resulting in significantly different estimates (also shown in Table S11).

10.13.2 Hudson ratio of averages

Authors: Gaurav Bhatia, Nick Patterson, Alkes L. Price

We computed F_{ST} estimates using the definition of Hudson et al.[94], equivalent to an earlier definition of Nei[95]. Estimates were computed using a ratio of averages, as opposed to an average of ratios[96]. Results are displayed in Table S11. In sequencing data with many rare variants, discrepancy between approaches for calculating F_{ST} can become very large, as rare population-private SNPs with low F_{ST} will cause large decreases in F_{ST}. We recommend that estimates of F_{ST} be computed using the ratio of averages approach[96], as this approach is robust to the inclusion of rare variants.

10.14 Quantifying potentially functional variants in the Phase 1 dataset

For the analysis presented in Table 2, we calculated the average number per individual of variant sites belonging to different potentially functional classes (detailed in the following sections), based on sites discovered in the low coverage genome-wide sequence data. These numbers are expressed as the mean per population, and the population range is given. Numbers are broken down into rare (<0.5%), low frequency (0.5-5%), and common (>5%) according to global derived allele frequency in the Phase 1 samples.

10.14.1 Coding variant classes

Authors: Yali Xue, Yuan Chen, Suganthi Balasubramanian, Lukas Habegger, Mark Gerstein, Chris Tyler-Smith

The coding variants included synonymous, nonsynonymous, stop-loss and indel-non-frameshift based on the annotation from GENCODE v7. Since variants may receive multiple annotations because of their consequences for alternative transcripts, a hierarchy stop-loss > nonsynonymous > synonymous was used, such that a variant was only counted once at its highest level in the hierarchy. Two sets of numbers were calculated: one for all sites, and a second for sites with GERP score >2, except for indels where GERP scores are not available and loss-of-function (LoF) variants which were considered damaging irrespective of the GERP score of the variant nucleotide(s). Potential LoF variants received additional annotation and curation, described below.

10.14.2 Identification and filtering of loss-of-function variants

Authors: Daniel MacArthur, Suganthi Balasubramanian, Mike Jin, Adam Frankish, Jennifer Harrow, Mark Gerstein, Chris Tyler-Smith

We examined SNV and short indel calls for variants predicted to result in the complete loss-of-function (LoF) of protein-coding genes. Annotations were performed using the Variant Annotation Tool[64].

LoF variants were defined as:

1. SNVs predicted to create a premature stop codon (stop_gained);
2. SNVs predicted to disrupt an essential splice site, i.e. variants found in the 2bp at either end of a spliced intron (splice_site);

3. Small insertions or deletions in a coding region that had a length that was not a multiple of 3, and were thus expected to disrupt the normal reading frame (frameshift_indel).

We next applied filters to these variants to remove likely sequencing and annotation artifacts, using similar approaches to those described in a recent analysis of LoF variants[97] that were identified as part of the 1000 Genomes pilot project. Variants satisfying any of the following criteria were regarded as likely artifactual:

1. Variants included in the phase 1 site list, but for which no individuals were explicitly called as carrying the non-reference allele (no_alt_calls);
2. Variants where the LoF allele was also inferred to be ancestral, based on comparison with non-human primate genomes (lof_anc);
3. Predicted truncating variants found in the first or last 5% of the coding sequence (near_start, near_stop);
4. Variants identified as likely artifacts through manual reannotation performed as part of the 1000 Genomes pilot project (pilot_filt);
5. Splice variants found in a noncanonical splice site, i.e. a site in the reference that did not follow the standard GT-AG rule (noncanonical);
6. Splice variants where the other splice site in that intron was noncanonical (other_noncanonical);
7. Variants found in an intron with a total length of less than 15 bp (short_intron);
8. Multiple indels in the same coding sequence with evidence for genetic linkage, which in combination would be expected to restore reading frame (linked_indel);

In addition to applying these criteria to all candidate LoF variants, we manually inspected the annotation evidence supporting all novel LoF variants with a global allele frequency above 5%, and removed those with poor transcript support, or for which other criteria supporting an annotation artifact were observed (manual_annot).

10.14.3 HGMD-DM and COSMIC SNPs

Authors: Yali Xue, Yuan Chen, Qasim Ayub, Edward V. Ball, Peter D. Stenson, David N. Cooper, and Chris Tyler-Smith

The set of SNPs overlapping between 1000 Genomes Phase 1 and the HGMD DM (Damaging Mutation) class (HGMD[98] version 2010.4) or the COSMIC[99] base substitutions (v56) were picked based on correspondence between the chromosome coordinate and allele. The overlapping sets of variants were then filtered by GERP score >2.

10.14.4 Non-coding variant classes

Authors: Xinmeng Jasmine Mu, Ekta Khurana, Yali Xue, Yuan Chen, Arif Harmanci, Cristina Sisu, Chris Tyler-Smith, and Mark Gerstein

The following non-coding variant categories were included: UTRs (5' and 3'), non-coding RNAs (lincRNA, miRNA, miscRNA, rRNA, snoRNA, snRNA), motif gain in TF peak, and motif loss in TF peak. The categories are based on the non-coding annotations derived from the GENCODE[65] v7 and ENCODE Integrative paper release[66,67]. The counts are redundant - i.e. no hierarchy was used and a variant was counted multiple times if it received multiple annotations.

The motif gain and loss analysis investigated SNPs that fall into TF-binding motifs for the 121 TFs assayed by ChIP-seq experiments in the ENCODE project. Data from all cell types were used for this analysis. The motif gain variant category is defined as a SNP whose derived allele has a higher frequency in the Position Weight Matrix (PWM) of the bound motif than the ancestral allele; the motif loss variant category is defined as a SNP whose derived allele has a lower frequency in the PWM of the bound motif than the ancestral allele.

All non-coding variant numbers were calculated for all variants and for variants with GERP >2.

10.14.5 Other conserved variants

Authors: Yali Xue, Yuan Chen, Xinmeng Jasmine Mu, Ekta Khurana, Mark Gerstein, and Chris Tyler-Smith

All sites with GERP score >2 were considered as 'Total conserved' sites, and all sites with GERP score > 2 but not in the above categories as 'Other conserved' sites.

10.15 Rare variant proportion for variants with functional consequences

Author: Tuuli Lappalainen

Figure 4a shows the correlation between evolutionary conservation and allele frequency in different functional annotation categories. This analysis was done on SNPs found in the low coverage data to have comparable data from coding and non-coding variants. Conservation of the SNP sites was measured by the mammalian GERP score[100]. Derived allele frequency was calculated across the entire sample, and the plot shows the proportion of sites with frequency < 0.5%. Sites without GERP score or ancestral allele information were excluded. The functional annotation is described in Section 8.1; additionally, the null category consists of 2 million random low coverage SNP sites that do not belong to any annotation category. In the figure, the lines represent medians of bins of size adjusted according to the number of SNPs in each annotation category, with 75% overlap between adjacent bins. The crosses adjacent to the axes show the overall medians of the annotation categories.

10.16 Estimation of excess nonsynonymous variants in KEGG pathways

Author: Adam Auton

Gene pathways were obtained from the KEGG database[101]. We selected pathways with more than 5 genes, giving a total of 186 pathways. For each pathway, we estimated the number nonsynonymous (NSyn) SNPs to synonymous (Syn) SNPs, separately for SNPs with minor allele frequency > 0.5% and < 0.5%. For each pathway, we estimated the excess of NSyn SNPs at low frequencies by calculating:

$$\text{Excess rare NSyn} = \#\text{NSyn}_{\text{MAF}<0.5\%} - (\#\text{Syn}_{\text{MAF}<0.5\%} \text{ x } \#\text{NSyn}_{\text{MAF}>0.5\%} / \#\text{Syn}_{\text{MAF}>0.5\%})$$

This formula provides an estimate of the excess number of rare nonsynonymous SNPs on the basis of the number expected from the NSyn/Syn ratio of common SNPs. The resulting estimates are compared to the conservation GERP score[100] in Figure S11 and Table S13.

10.17 Nucleotide diversity around CTCF binding motifs

Author: Adam Auton

We identified all copies of a putative CTCF-binding motif CCMYCTNNNGG, which was selected on the basis of consensus between previous studies[102,103]. In total, we identified 386,528 such motifs. Of these, we identified 17,970 motifs that intersected with peaks in the ENCODE CTCF-binding annotation in the GM12878 cell line, obtained from the UCSC genome browser (CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington, peaks replicate 1, although results are similar using a different replicate)[53,104]. For each motif in a peak, we tried to identify an identical motif outside of a peak (without replacement). This could be achieved for 17,485 motifs (97.3%).

We identified the SNPs at every base in each motif, and summed across all motifs, with the resulting plot shown in Figure 4c, partitioned by SNP frequency. Motifs contained within CTCF-bound regions show lower levels of polymorphism than those outside of these regions, and the SNPs within CTCF-bound motifs tend to be less common than those outside binding sites, indicating an increased level of constraint on CTCF-bound loci.

10.18 Population differentiation of functional SNPs

Authors: Adam Auton and Gil McVean

To produce Figure S13, we first partitioned variants on the basis of their allele count within each population. For each allele-count within each population, we estimated F_{ST} to all other populations using just nonsynonymous variants, and separately using just synonymous variants. We then plotted the proportion of estimates with a larger nonsynonymous estimate as a function of allele count in the target population. Although estimates of F_{ST} are sensitive to allele frequency, we do not expect nonsynonymous and synonymous sites (which are also interleaved) to show systematic differences in differentiation unless selection is operating differently on the different variant types. The systematically higher differentiation of nonsynonymous variants at low frequencies is therefore consistent with the effects of purifying selection. At higher frequencies the difference between nonsynonymous and synonymous variants decreases, perhaps because higher-frequency nonsynonymous variants are less likely to be under strong purifying selection than rarer ones.

10.19 Variants in linkage disequilibrium with focal GWAS SNPs

Author: Hyun Min Kang

The average number of variants in linkage disequilibrium to focal SNPs identified in GWAS for each SNP in the GWAS catalog (as of May 16, 2012)[105], the |D'| and r^2 with 1000G variants within 1Mb window on each side was evaluated within each continental population or across all the individuals. For each distance threshold bin, the number of SNPs with r^2 >=.5 (or |D'|=1) beyond the distance was counted, and averaged across all GWAS catalog SNPs. Control SNPs are selected among the SNPs assayed in Affymetrix 500k SNP arrays to account for potential ascertainment bias, and matched by European allele frequency, and the distance to the nearest gene. Trans-ethnic fine mapping was evaluated by taking the minimum r^2 or |D'| across the four continental populations. The counts are refined by different categories, such as HapMap 2+3 variants only, variants found in the 1000G pilot study, all 1000G variants, non-synonymous coding variants, variants with GERP conservation score 2 or greater, and variant type. The results are shown in Figure 5b and Table S15.

10.20 Comparison of 1000 Genomes Phase 1 to UK10K study

Author: Klaudia Walter

The UK10K project, a collaboration between the Wellcome Trust Sanger Institute and multiple research centres in the UK and Finland, aims to research the relationship between genetic variants of a broad frequency range and measures of health and disease status in a variety of study designs using 4,000 whole low-coverage genomes and 6,000 high-coverage exomes.

The December 2011 UK10K release (REL-2011-12-01) contains 2432 low-coverage genomes. The percentage of variable sites shared between this UK10K release and the Phase 1 release of the 1000 Genomes Project is 46.0% at a minor allele frequency (MAF) of 0.1%, 95.8% at 1% and 97.7% at 5%. The MAF was estimated from the allele counts in the UK10K data, and the percentage was computed for each reported MAF allowing a small range of +/- 10%.

10.21 Comparison of 1000 Genomes Phase 1 to SardiNIA study

Author: Goncalo Abecasis

The SardiNIA Medical Sequencing Discovery Project is a study of the genetics of age-related traits including blood lipid levels and personality in a Sardinian population cohort. As of June 2012, the genomes of 2120 individuals had been sequenced to 3.5X coverage by this project at the CSR4 research institute, in Pula, Sardinia, and at the University of Michigan. Samples have been sequenced using Illumina GAII and HiSeq 2000 instruments with 100 - 120 bp paired-end reads.

We performed a comparison of the SardiNIA dataset to the 1000 Genomes Project Phase 1 release. Sites within the 'population-genetics mask' (see Section 10.4: chosen so as to minimize the effects of between-project differences in variant detection algorithms) with variants at minor allele frequency of 0.1% in the SardiNIA study were also variable in the 1000 Genomes dataset 23.7% of the time, rising to 76.9% and 99.3% for variants with MAF of 1.0% and 5.0% respectively.

11 Accessing 1000 Genomes data

Authors: Laura Clarke, Xiangqun Zheng-Bradley, and Richard E. Smith

A full description of data management and community access can be found in Clarke *et al.*[106] The 1000 Genomes Project has two mirrored FTP sites that follow the same basic structure.

- Europe: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
- USA: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/

A description of the FTP structure can be found in the README file contained in the top-level directory.

Tutorials explaining recommended methods for accessing and using the data have been made available at: http://www.1000genomes.org/using-1000-genomes-data

Finally, support for using the 1000 Genomes Project data can be obtained via email: info@1000genomes.org

12 References

1 Coriell Institute. *NHGRI Sample Repository for Human Genetic Research*, <http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11> (2012).
2 Coriell Institute. *Guidelines for Referring to the Populations in Publications and Presentations*, <http://ccr.coriell.org/Sections/Support/NHGRI/NHGRI_Pop_Ref.aspx?PgId=688> (2012).
3 Freshney, R. I. & Freshney, M. G. *Culture of immortalized cells*. (Wiley-Liss, 1996).

4 Coriell Institute. *Frequently Asked Questions about Lymphoblastoid Cell Cultures* <http://ccr.coriell.org/Sections/Support/Global/Lymphoblastoid.aspx?PgId=213> (2012).

5 Coriell Institute. *Genotyping with Microsatellites Assures Cell Line Identity and Culture Purity*, <http://ccr.coriell.org/Sections/Support/Global/QCgenotype.aspx?PgId=412> (2012).

6 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

7 Shumway, M., Cochrane, G. & Sugawara, H. Archiving next generation sequencing data. *Nucleic acids research* **38**, D870-871, doi:10.1093/nar/gkp1078 (2010).

8 Wysoker, A. *Picard*, <http://picard.sourceforge.net/> (2012).

9 Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology* **12**, R1, doi:10.1186/gb-2011-12-1-r1 (2011).

10 Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature methods* **5**, 1005-1010, doi:10.1038/nmeth.1270 (2008).

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

12 The 1000 Genomes Project Consortium. *Exome Pull Down*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/exome_pull_down> (2012).

13 Human Genome Sequencing Center - Baylor College of Medicine. *Preparation of SOLiD Capture Libraries*, <http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf> (2012).

14 Jostins, L., Danecek, P., Li, H. & Durbin, R. *GLFTools*, <http://sourceforge.net/projects/samtools/files/glftools> (2009).

15 Kang, H. M. *et al.* *VerifyBamID*, <http://genome.sph.umich.edu/wiki/VerifyBamID> (2011).

16 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).

17 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

18 Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome research* **11**, 1725-1729, doi:10.1101/gr.194201 (2001).

19 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

20 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:gr.107524.110 [pii]

10.1101/gr.107524.110 (2010).

21 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

22 Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157-1158, doi:10.1093/bioinformatics/btr076 (2011).

23 Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4**, e7767, doi:10.1371/journal.pone.0007767 (2009).

24 Li, W., Stromberg, M. P. & Marth, G. *MOSAIK*, <https://github.com/wanpinglee/MOSAIK> (2012).

25 Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-1692, doi:10.1093/bioinformatics/btr174 (2011).

26 The 1000 Genomes Project Consortium. *Broad Exome Illumina BAM files*, <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/technical/other_exome_alignments/> (2012).

27 Wang, Y. & Yu, F. *SNPTools*, <http://www.hgsc.bcm.tmc.edu/cascade-tech-software_snp_tools-ti.hgsc> (2011).

28 Wang, Y., Lu, J., Yu, J., Gibbs, R. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *In revision* (2012).

29 Kang, H. M. *et al. UMAKE*, <http://genome.sph.umich.edu/wiki/UMAKE> (2012).

30 Brent, R. P. *Algorithms for minimization without derivatives*. (Prentice-Hall, 1973).

31 Consortium., T. G. P. *Consensus Target Capture Region*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/exome_pull_down/)> (2012).

32 Challis, D. *et al.* An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC bioinformatics* **13**, 8, doi:10.1186/1471-2105-13-8 (2012).

33 Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome research* **20**, 273-280, doi:10.1101/gr.096388.109 (2010).

34 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

35 Garrison, E. K. *Freebayes*, <https://github.com/ekg/freebayes> (2012).

36 Garrison, E. K. *ogap - a gap opening read aligner for BAM data streams*, <https://github.com/ekg/ogap> (2012).

37 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

38 Garrison, E. K. *vcflib - a simple C++ library for parsing and manipulating VCF files*, <https://github.com/ekg/vcflib> (2012).

39 Ward, A. *vcfCTools - a C++ implementation of vcfPytools*, <https://github.com/AlistairNWard/vcfCTools> (2012).

40 Durbin, R. *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. (Cambridge University Press, 1998).

41 Albers, C. A. *et al.* Dindel: accurate indel calls from short-read data. *Genome research* **21**, 961-973, doi:10.1101/gr.112326.110 (2011).

42 Rimmer, A., Mathieson, I., Lunter, G. & McVean, G. *Platypus: An Integrated Variant Caller* <http://www.well.ox.ac.uk/platypus> (2012).

43 Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**, 830-839, doi:10.1101/gr.115907.110 (2011).

44 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

45 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* **21**, 974-984, doi:10.1101/gr.114876.110 (2011).

46 Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. . *Bioinformatics* **In Press** (2012).

47 Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-426, doi:10.1126/science.1149504 (2007).

48 Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-276, doi:10.1038/ng.768 (2011).

49 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

50 Fan, X., Chen, K. & Chen, L. *TIGRA-SV*, <http://gmt.genome.wustl.edu/tigra-sv/current/> (2012).

51 Green, P. *Phred, Phrap, Consed*, <http://www.phrap.org/phredphrapconsed.html#block_phrap> (1996).

52 Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **27**, 595-603, doi:10.1093/bioinformatics/btq713 (2011).

53 Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39**, D876-882, doi:10.1093/nar/gkq963 (2011).

54 Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**, 47-55, doi:nbt.1600 [pii]
10.1038/nbt.1600 (2010).

55 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-223, doi:10.1016/j.ajhg.2009.01.005 (2009).

56 Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* **21**, 940-951, doi:10.1101/gr.117259.110 (2011).

57 Huang, L. *et al.* Genotype-imputation accuracy across worldwide human populations. *American journal of human genetics* **84**, 235-250, doi:10.1016/j.ajhg.2009.01.013 (2009).

58 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-470, doi:10.1534/g3.111.001198 (2011).

59 Karafet, T. M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome research* **18**, 830-838, doi:10.1101/gr.7172008 (2008).

60 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81, doi:10.1126/science.1181498 (2010).

61 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

62 Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics* **23**, 147, doi:10.1038/13779 (1999).

63 The 1000 Genomes Project Consortium. *GENCODE7 gene annotation model*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/annotation_sets/> (2012).

64 Gerstein Lab. *Variant Annotation Tool (VAT)*, <http://vat.gersteinlab.org/> (2012).

65 Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome research* **In Press** (2012).

66 The ENCODE Project Consortium. Initial Analysis of the Encyclopedia of DNA Elements in the Human Genome. *Nature* **In Press** (2012).

67 Gerstein, M. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **In Press** (2012).

68 Yip, K. Y. *et al.* Classification of genomic regions based on experimentally-determined binding sites of more than 120 transcription-related factors in the whole human genome. *Genome biology* **In Press** (2012).

69 Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research* **18**, 1829-1843, doi:10.1101/gr.076521.108 (2008).

70 The 1000 Genomes Project Consortium. *Ancestral allele reference sequences*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/> (2012).

71 Carneiro, M. O. *et al.* Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics* (2012).

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

73 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

74 The 1000 Genomes Project Consortium. *Validation results*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/snps/> (2012).
75 The 1000 Genomes Project Consortium. *Integrated Call Sets*, 2012).
76 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
77 Exome Chip Design Consortium. *Exome Chip Design*, <http://genome.sph.umich.edu/wiki/Exome_Chip_Design> (2012).
78 Kent, J. *In-Silico PCR*, <http://genome.ucsc.edu/cgi-bin/hgPcr> (2012).
79 The 1000 Genomes Project Consortium. *Accessible Genome Masks*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/> (2012).
80 The 1000 Genomes Project Consortium. *Sample Pedigree*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/integrated_call_samples.20101123.ped > (2012).
81 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).
82 The 1000 Genomes Project Consortium. *OMNI SHAPEIT haplotypes*, <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/omni_haplotypes/ > (2012).
83 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).
84 Ramantani, G. *et al.* Expanding the phenotypic spectrum of lupus erythematosus in Aicardi-Goutieres syndrome. *Arthritis and rheumatism* **62**, 1469-1477, doi:10.1002/art.27367 (2010).
85 Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **In press** (2012).
86 Mao, X. *et al.* A genomewide admixture mapping panel for Hispanic/Latino populations. *American journal of human genetics* **80**, 1171-1178, doi:10.1086/518564 (2007).
87 Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359-1367, doi:10.1093/bioinformatics/bts144 (2012).
88 Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5**, e1000519, doi:10.1371/journal.pgen.1000519 (2009).
89 Churchhouse, C. & Marchini, J. Multi-way admixture deconvolution using phased or unphased ancestral panels. *Genet. Epi.* **In press** (2012).
90 The 1000 Genomes Project Consortium. *Local Ancestry Inference*, <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/ancestry_deconvolution/> (2012).
91 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

92 Weir, B. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution*, 1358-1370 (1984).

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

94 Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589 (1992).

95 Nei, M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**, 3321-3323 (1973).

96 Ding, Z. L. *et al.* Origins of domestic dog in southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity* **108**, 507-514, doi:10.1038/hdy.2011.114 (2012).

97 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

98 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13, doi:10.1186/gm13 (2009).

99 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-950, doi:10.1093/nar/gkq929 (2011).

100 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

101 Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-114, doi:10.1093/nar/gkr988 (2012).

102 Essien, K. *et al.* CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome biology* **10**, R131, doi:10.1186/gb-2009-10-11-r131 (2009).

103 Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* **4**, e1000138, doi:10.1371/journal.pgen.1000138 (2008).

104 Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046, doi:10.1371/journal.pbio.1001046 (2011).

105 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

106 Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nature methods* **9**, 459-462, doi:10.1038/nmeth.1974 (2012).

Table S1 Low-coverage sequence coverage

Population	Platform	Sample Number	Total Raw Base Pairs	Total Base Pairs	Mapped Base Pairs	Estimated Mean Coverage (*)
ASW	ILLUMINA	50	1,089,813,717,972	968,854,526,076	860,327,970,070	6.07
ASW	SOLID	11	500,286,849,800	342,449,243,700	216,331,285,800	6.94
ASW	all	61	1,590,100,567,772	1,311,303,769,776	1,076,659,255,870	6.22
CEU	ILLUMINA	79	1,502,960,787,784	1,294,604,937,320	1,094,649,530,588	4.89
CEU	LS454	13	146,488,692,190	131,480,819,959	98,961,124,472	2.68
CEU	all	85	1,649,449,479,974	1,426,085,757,279	1,193,610,655,060	4.95
CHB	ILLUMINA	81	1,265,574,070,165	1,134,008,661,926	984,021,701,292	4.28
CHB	SOLID	16	630,769,366,650	417,614,608,100	244,511,888,650	5.39
CHB	all	97	1,896,343,436,815	1,551,623,270,026	1,228,533,589,942	4.47
CHS	ILLUMINA	92	1,418,189,794,341	1,299,646,744,020	1,071,501,746,928	4.11
CHS	SOLID	8	200,584,536,300	128,336,437,000	84,922,544,700	3.74
CHS	all	100	1,618,774,330,641	1,427,983,181,020	1,156,424,291,628	4.08
CLM	ILLUMINA	50	1,081,008,434,634	997,880,714,716	809,514,266,283	5.71
CLM	SOLID	10	499,995,107,550	326,310,219,800	205,501,075,700	7.25
CLM	all	60	1,581,003,542,184	1,324,190,934,516	1,015,015,341,983	5.97
FIN	ILLUMINA	75	1,138,524,186,998	1,044,187,677,032	836,323,008,056	3.93
FIN	SOLID	18	608,575,692,700	445,003,405,050	293,354,088,000	5.75
FIN	all	93	1,747,099,879,698	1,489,191,082,082	1,129,677,096,056	4.28
GBR	ILLUMINA	70	1,215,027,872,642	1,101,288,886,042	824,982,280,523	4.16
GBR	SOLID	19	746,168,251,800	525,113,156,350	340,810,710,650	6.33
GBR	all	89	1,961,196,124,442	1,626,402,042,392	1,165,792,991,173	4.62
IBS	ILLUMINA	6	91,768,475,800	86,301,704,300	79,602,973,565	4.68
IBS	SOLID	8	365,162,926,050	240,650,750,400	160,414,631,900	7.07
IBS	all	14	456,931,401,850	326,952,454,700	240,017,605,465	6.05
JPT	ILLUMINA	78	1,934,078,557,132	1,700,345,329,015	1,396,406,178,908	6.31
JPT	SOLID	11	469,831,458,750	312,859,319,500	180,345,872,950	5.78
JPT	all	89	2,403,910,015,882	2,013,204,648,515	1,576,752,051,858	6.25
LWK	ILLUMINA	83	1,618,796,704,250	1,488,826,665,046	1,308,949,191,963	5.56
LWK	SOLID	14	632,314,850,250	430,833,584,700	236,644,976,100	5.96
LWK	all	97	2,251,111,554,500	1,919,660,249,746	1,545,594,168,063	5.62
MXL	ILLUMINA	54	1,092,715,400,556	1,013,311,530,884	863,106,820,709	5.64
MXL	SOLID	12	541,920,312,100	372,689,964,700	232,992,623,250	6.85
MXL	all	66	1,634,635,712,656	1,386,001,495,584	1,096,099,443,959	5.86
PUR	ILLUMINA	52	1,201,929,599,317	1,089,276,083,789	807,903,247,949	5.48
PUR	SOLID	3	54,314,001,150	40,579,894,850	29,324,996,900	3.45
PUR	all	55	1,256,243,600,467	1,129,855,978,639	837,228,244,849	5.37
TSI	ILLUMINA	98	1,432,811,860,807	1,333,997,141,749	1,234,369,684,334	4.44
TSI	all	98	1,432,811,860,807	1,333,997,141,749	1,234,369,684,334	4.44
YRI	ILLUMINA	76	1,422,168,538,822	1,237,450,311,835	1,060,031,524,163	4.92
YRI	SOLID	12	534,037,533,950	349,126,967,950	193,688,013,900	5.69
YRI	all	88	1,956,206,072,772	1,586,577,279,785	1,253,719,538,063	5.02
Total	ILLUMINA	944	17,505,368,001,220	15,789,980,913,750	13,231,690,125,331	4.94
Total	LS454	13	146,488,692,190	131,480,819,959	98,961,124,472	2.68
Total	SOLID	142	5,783,960,887,050	3,931,567,552,100	2,418,842,708,500	6.01
Total	**all**	**1092**	**23,435,817,580,460**	**19,853,029,285,809**	**15,749,493,958,303**	**5.09**

* Assuming an accessible genome of 2.84Gb

Table S2 Exome sequence coverage

Population	Platform	Sample Number	Total Base Pairs	Mapped Base Pairs	Mapped to Target	Estimated Mean Coverage in Target (*)
ASW	ILLUMINA	49	481,145,063,324	364,701,630,723	104,015,041,744	72.14
ASW	SOLID	9	132,601,206,696	98,844,286,498	12,777,136,555	48.25
ASW	all	58	613,746,270,020	463,545,917,221	116,792,178,299	68.43
CEU	ILLUMINA	55	781,710,583,271	611,783,336,621	162,444,963,408	100.37
CEU	SOLID	26	548,672,944,700	382,456,472,127	50,535,591,941	66.05
CEU	all	81	1,330,383,527,971	994,239,808,748	212,980,555,349	89.36
CHB	ILLUMINA	70	811,460,564,028	656,573,012,363	208,394,101,425	101.17
CHB	SOLID	23	381,992,713,440	280,204,262,734	37,342,955,772	55.18
CHB	all	93	1,193,453,277,468	936,777,275,097	245,737,057,197	89.79
CHS	ILLUMINA	90	857,138,087,206	714,503,886,102	251,302,974,303	94.89
CHS	SOLID	10	156,047,716,700	114,921,715,511	14,521,688,209	49.35
CHS	all	100	1,013,185,803,906	829,425,601,613	265,824,662,512	90.34
CLM	ILLUMINA	41	431,647,861,041	335,038,858,436	93,040,137,384	77.12
CLM	SOLID	19	297,290,521,550	213,373,947,718	27,039,889,718	48.36
CLM	all	60	728,938,382,591	548,412,806,154	120,080,027,102	68.01
FIN	ILLUMINA	75	716,066,832,301	584,073,449,816	199,492,310,600	90.39
FIN	SOLID	17	270,302,462,280	199,112,568,168	25,641,811,876	51.26
FIN	all	92	986,369,294,581	783,186,017,984	225,134,122,476	83.16
GBR	ILLUMINA	60	771,096,882,996	602,232,267,940	169,836,973,098	96.19
GBR	SOLID	26	443,801,705,735	318,174,886,647	44,696,071,194	58.42
GBR	all	86	1,214,898,588,731	920,407,154,587	214,533,044,292	84.77
JPT	ILLUMINA	68	964,872,725,317	748,071,660,429	237,137,258,342	118.51
JPT	SOLID	19	376,101,495,295	274,988,883,452	33,233,029,976	59.44
JPT	all	87	1,340,974,220,612	1,023,060,543,881	270,370,288,318	105.61
LWK	ILLUMINA	24	303,943,651,024	228,078,920,282	74,710,024,695	105.79
LWK	SOLID	62	987,305,907,096	740,239,916,698	91,749,291,260	50.29
LWK	all	86	1,291,249,558,120	968,318,836,980	166,459,315,955	65.78
MXL	ILLUMINA	54	572,707,204,495	429,456,560,453	114,487,895,932	72.05
MXL	SOLID	11	164,258,975,400	123,917,538,096	15,551,396,516	48.04
MXL	all	65	736,966,179,895	553,374,098,549	130,039,292,448	67.99
PUR	ILLUMINA	48	890,850,110,179	679,446,353,545	161,346,605,090	114.23
PUR	all	48	890,850,110,179	679,446,353,545	161,346,605,090	114.23
TSI	ILLUMINA	60	701,190,245,690	547,862,286,112	157,297,411,348	89.09
TSI	SOLID	35	585,031,422,890	413,598,481,830	62,757,299,612	60.93
TSI	all	95	1,286,221,668,580	961,460,767,942	220,054,710,960	78.72
YRI	ILLUMINA	71	1,066,578,885,883	826,820,770,280	243,135,647,582	116.37
YRI	SOLID	17	304,701,577,167	220,513,527,517	26,860,376,444	53.69
YRI	all	88	1,371,280,463,050	1,047,334,297,797	269,996,024,026	104.27
Total	ILLUMINA	765	9,350,408,696,755	7,328,642,993,102	2,176,641,344,951	96.69
Total	SOLID	274	4,648,108,648,949	3,392,004,935,418	442,706,539,073	54.91
Total	**all**	**1039**	**13,998,517,345,704**	**10,720,647,928,520**	**2,619,347,884,024**	**85.67**

* Assuming a exome target size of 29.4Mb

Table S3 Samples with OMNI 2.5M genotypes available, phased using family data available

POP	# duos	# trios	# unrelateds	Total samples
ACB	2	21	31	98
ASW	20	12	21	97
CDX	0	0	100	100
CEU	0	2	98	104
CHB	0	0	100	100
CHD	0	0	1	1
CHS	0	50	0	150
CLM	1	34	3	107
FIN	0	0	100	100
GBR	1	0	99	101
GIH	0	0	93	93
IBS	1	48	1	147
JPT	0	0	100	100
KHV	2	19	57	118
LWK	1	0	98	100
MKK	0	0	31	31
MXL	1	29	11	100
PEL	0	34	2	104
PUR	0	35	6	111
TSI	0	0	100	100
YRI	13	43	6	161
Total	**84**	**981**	**1058**	**2123**

Note this includes samples beyond Phase1 (and not all Phase 1 samples)
Haplotypes can be found at
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/omni_haplotypes/

Table S4 Low-coverage SNP validation

	Total	True SNP	False SNP	No call	FDR (%)	No call rate (%)
Total	287	276	5	6	1.8	2.1
Singletons	70	65	3	2	4.4	2.9
MAF<0.01	134	131	2	1	1.5	0.7
0.01<MAF<0.05	33	33	0	0	0	0
MAF>0.05	50	47	0	3	0	6

A total of 287 SNP sites were included in the final SNP validation results. True and false SNPs are those confirmed or rejected by the consensus of the four validation experiments. "No call" SNPs did not produce a reliable result in any of the validation experiments. The false discovery rate (FDR) is calculated by dividing the number of false SNPs by the sum of the true and false SNPs. The no call rate is the no call SNPs divided by the total SNPs. The data has also been split by minor allele frequency (MAF). The MAF<0.01 category does not include singleton SNPs.

Results for each SNP can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/snps/

Table S5 Exome SNP validation

Novel exome consensus SNP validation

	Total	True SNP	False SNP	No call	FDR (%)	No call rate (%)
Total	200	185	3	12	1.6	6
singleton	100	92	1	7	1.1	7
AF < 1%	50	47	2	1	4.1	2
AF > 1%	50	46	0	4	0	8
Novel	100	84	2	14	2.3	14

Center-unique exome SNP validation

	In consensus	Not in consensus	Total validated	True SNP	False SNP	No call	FDR (%)	No call rate (%)
Illumina								
BC Unique	28	74	20	2	4	14	67%	70
BCM Unique	77	157	20	13	5	2	28%	10
UM Unique	175	74	20	13	1	6	7%	30
SOLiD								
BC Unique	63	28	17	6	7	4	54%	23.5
BCM Unique	238	200	20	4	14	2	78%	10
UM Unique	83	117	20	0	16	4	100%	20

A total of 417 SNP sites were included in three exome SNP validation experiments. Table S5a shows the validation results of exome consensus SNPs stratified by allele frequency and novelty. The AF<0.01 category does not include singleton SNPs. SNPs are considered novel if they are not found in the low coverage SNP call set or in dbSNP135. Table S5b shows the results of different centers' unique exome SNP calls that were not included in the exome consensus set. In both tables, true and false SNPs are those confirmed or rejected by PCR-Roche 454 validation. "No call" SNPs did not produce a reliable result in the validation experiment. The false discovery rate (FDR) is calculated by dividing the false SNPs by the sum of the true and false SNPs. The no call rate is the no call SNPs divided by the total SNPs.

Results for each SNP can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/snps/

Table S6. Low-coverage INDEL validation summary

	Total	True INDEL	False INDEL	No call	FDR (%)	No call rate (%)	AFFY-FDR-BEFORE-SVM	AFFY-FDR-AFTER-SVM
Total	93	49	27	17	35.5	18.3	12.5	5.4
MAF<0.01	15	4	10	1	71.4	7.1	13.8	8.1
0.01<MAF<0.10	36	22	6	8	27.3	22.2	12.1	5.2
MAF>0.10	42	23	11	8	32.4	19	12.2	3.7

A total of 93 INDEL sites were included in the INDEL validation study. True and false SNPs are those confirmed or rejected by the consensus of the three validation experiments. “No call” SNPs did not produce a reliable result in any of the validation experiments (some were not amplified by PCR, others did not produce reliable sequencing calls). The false discovery rate (FDR) is calculated by dividing the number of false INDELs by the sum of the true and false INDELs. The no call rate is the number of no call INDELs divided by the total number of INDELs. AFFY-FDR-BEFORE-SVM and AFFY-FDR-AFTER-SVM are the estimated false discovery rate before and after applying SVM filtering calculated as a proportion of monomorphic sites genotyped in Affymetrix Exome Array. The data has also been split by minor allele frequency (MAF).

Individual results for each indel can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/indels/

A list of indel sites excluded in the post-hoc filtering can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/excluded_indel_sites/

Table S7 SV call sets, estimated False Discovery Rate (FDR), and number of sites/samples evaluated

Algorithm	Variants called after merging	Estimated per algorithm FDR from initial validation			Selection criterion for promotion to discovery set	Number in discovery set	Inferred FDR of discovery set
		Validation Method					
		Omni 2.5	PCR	Array CGH			
BreakDancer	20,388	14.1%	12.0%	13.9%	calls validated by Omni 2.5	5,914	1% (assumed)
		n=6,959	n=75	n=11,417	(p < 0.01)		
CNVnator	20,062	74.1%	29.6% [1]	38.2%	calls validated by Omni 2.5	2,084	1% (assumed)
		n=5,097	n=27	n=58,293	(p < 0.01)		
Delly	38,758	57.2%	0.0% [2]	15.9%	calls validated by Omni 2.5	5,073	1% (assumed)
		n=10,822	n=78	n=4,092	(p < 0.01)		
Genome STRiP	18,912	1.5%	2.9%	4.2%	all calls	18,912	1.5% - 4.2% across validation methods
		n=10,386	n=70	n=12,187			
Pindel	41,370	83.0%	40.0% [1]	47.9%	calls validated by Omni 2.5	1,294	1% (assumed)
		n=6,619	n=10	n=57,504	(p < 0.01)		
Non-redundant total	113,649					23,594	1.4% - 3.7%
Genotyped set [3]						14,422	1.4% - 3.7%

Summary of implied FDR of discovery set after construction [4]

	Experimental method			
	aCGH	PCR	Both aCGH and PCR	Union aCGH and PCR
Sites attempted	3,305	87	98	3,490
Sites validated	3,186	64	93	3,343
Sites invalidated	70	2	0	72
Sites inconclusive or discordant	49	21	5	75
Estimated FDR	2.1%	3.0%	0.0%	2.1%

Notes

[1] For CNVnator and Pindel, the calls for PCR validation were originally selected from a more sensitive superset.
The reported FDR is based on a more stringent subset used in the other validation experiments and subsequent analyses.

[2] For Delly, the calls for PCR validation were selected on a per-observation (frequency weighted) basis sampled disproportionately from sites with high deletion allele frequency.
The resulting FDR estimate is not comparable to the other call sets where calls were selected on a per-site basis, independent of deletion allele frequency.

[3] Sites were genotyped if sufficient data was available to calculate accurate genotype likelihoods; additional non-polymorphic and redundant sites were removed during genotyping.

[4] Array CGH and PCR validation sites were selected randomly from the calls in the individual call sets from each method, not the promoted discovery set.

The array CGH and PCR validation results were not available when the promoted discovery set was created.
Omni 2.5 validation uses SNP array probe intensities and is more sensitive for longer deletion events.
Array CGH validation was performed in 25 selected samples and only tests sites discovered in those samples.
Merged discovery sets are available as a supplementary data file.

Results for each site can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/sv/

Table S8 LOF SNPs validation results

Frequency Class	Total	True SNP	False SNP	No Call	FDR (%)	No Call rate (%)
Singleton	1183	1078	34	71	3.1	6
Doubleton	150	129	8	13	5.8	8.7
Tripleton	35	25	3	7	10.7	20
<1% (3Q=0.4%)	88	46	17	25	27	28.4
1% - 5%	17	3	4	10	57.1	58.8
>5%	8	1	4	3	80	37.5
Total	**1481**	**1282**	**70**	**129**	**5.2**	**8.7**

A total of 1,481 SNP sites were included in LOF PCR-Roche 454 validation. True and false SNPs are those confirmed or rejected by PCR-Roche 454 validation. "No call" SNPs did not produce a reliable result. The false discovery rate (FDR) is calculated by dividing the false SNPs by the sum of the true and false SNPs. The no call rate is the no call SNPs divided by the total SNPs. The data has been split by allele frequency (AF). The AF<0.01 category does not include singleton, doubleton and tripleton SNPs.

Results for each variant can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/snps/

Table S9 Formation mechanisms of large deletions

Mechanism	< 500 bp	500 - 1000 bp	1 kb - 10 kb	10 kb +	All
NAHR	9 (2.6%)	294 (23.3%)	1420 (22.6%)	255 (24.7%)	1978 (22.1%)
NHR	284 (82.8%)	889 (70.4%)	4642 (73.7%)	748 (72.4%)	6563 (73.5%)
MEI	47 (13.7%)	67 (5.3%)	124 (2.0%)	0 (0%)	238 (2.7%)
VNTR	2 (0.6%)	7 (0.6%)	64 (1.0%)	23 (2.2%)	96 (1.1%)
Undefined	1 (0.3%)	6 (0.5%)	45 (0.7%)	7 (0.7%)	59 (0.7%)
Total	343 (100%)	1263 (100%)	6295 (100%)	1033 (100%)	8934 (100%)

NAHR: Non-allelic homologous recombination

NHR: non-homologous rearrangements (including non-homologous end-joining and microhomology-mediated break-induced replication)

VNTR: variable number of tandem repeats

MEI: mobile element insertion

Table S10 Cryptic relationships identified by genome-wide SNP analysis

Population	Sample 1	Sample 2	Relationship	IBD0	IBD1	IBD2
ASW	NA19713	NA19985	Sibling	0.3	0.51	0.19
ASW	NA20289	NA20341	Sibling	0.23	0.53	0.24
ASW	NA20334	NA20336	Sibling	0.24	0.51	0.25
ASW	NA19625	NA20414	Second-order	0.43	0.57	0
ASW	NA20359	NA20363	Second-order	0.52	0.48	0
MXL	NA19660	NA19685	Parent/Child	0	1	0
MXL	NA19661	NA19685	Parent/Child	0	1	0
MXL	NA19675	NA19678	Parent/Child	0	1	0
MXL	NA19675	NA19679	Parent/Child	0	1	0
MXL	NA19660	NA19672	Sibling	0.24	0.48	0.28
MXL	NA19657	NA19753	Second-order	0.47	0.51	0.02
MXL	NA19660	NA19664	Second-order	0.57	0.43	0
MXL	NA19664	NA19672	Second-order	0.46	0.54	0
MXL	NA19672	NA19685	Second-order	0.44	0.56	0
MXL	NA19726	NA19738	Second-order	0.49	0.51	0
CHS	HG00656	HG00702	Parent/Child	0	1	0
CHS	HG00657	HG00702	Parent/Child	0	1	0
CHS	HG00501	HG00512	Sibling	0.28	0.48	0.24
CHS	HG00501	HG00524	Sibling	0.24	0.5	0.26
CHS	HG00512	HG00524	Sibling	0.2	0.52	0.28
CHS	HG00577	HG00584	Sibling	0.21	0.55	0.24
CHS	HG00578	HG00581	Sibling	0.19	0.54	0.27
CHS	HG00578	HG00635	Sibling	0.22	0.55	0.23
CHS	HG00581	HG00635	Sibling	0.26	0.55	0.19
CHS	HG00418	HG00427	Second-order	0.49	0.51	0
LWK	NA19313	NA19331	Parent/Child	0	1	0
LWK	NA19381	NA19382	Parent/Child	0	1	0
LWK	NA19445	NA19453	Parent/Child	0	1	0
LWK	NA19469	NA19470	Parent/Child	0	1	0
LWK	NA19331	NA19334	Sibling	0.29	0.48	0.23
LWK	NA19347	NA19352	Sibling	0.23	0.52	0.25
LWK	NA19373	NA19374	Sibling	0.24	0.5	0.26
LWK	NA19396	NA19397	Sibling	0.23	0.53	0.24
LWK	NA19434	NA19444	Sibling	0.27	0.51	0.22
LWK	NA19443	NA19470	Sibling	0.26	0.49	0.25
LWK	NA19313	NA19334	Second-order	0.51	0.49	0
LWK	NA19380	NA19382	Second-order	0.43	0.57	0
LWK	NA19434	NA19453	Second-order	0.59	0.41	0
LWK	NA19443	NA19469	Second-order	0.54	0.46	0
LWK	NA19444	NA19453	Second-order	0.47	0.53	0

Additional information can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/cryptic_relation_analysis/

Table S11 Pairwise estimates of FST

Weir and Cockerham Estimator

	LWK	YRI	ASW	CLM	MXL	PUR	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT
LWK		0.56%	0.65%	4.09%	4.63%	3.77%	5.06%	5.24%	5.08%	4.69%	4.98%	5.78%	5.85%	5.78%
YRI	0.56%		0.57%	4.72%	5.30%	4.37%	5.79%	6.00%	5.81%	5.66%	5.69%	6.55%	6.64%	6.55%
ASW	0.65%	0.57%		2.93%	3.63%	2.54%	4.23%	4.47%	4.26%	2.82%	4.22%	5.46%	5.56%	5.41%
CLM	4.09%	4.72%	2.93%		0.51%	0.34%	0.98%	1.19%	0.99%	0.42%	0.98%	3.51%	3.63%	3.54%
MXL	4.63%	5.30%	3.63%	0.51%		0.99%	1.82%	1.86%	1.81%	1.69%	1.83%	3.10%	3.24%	3.15%
PUR	3.77%	4.37%	2.54%	0.34%	0.99%		0.90%	1.22%	0.93%	0.25%	0.85%	3.95%	4.06%	3.98%
CEU	5.06%	5.79%	4.23%	0.98%	1.82%	0.90%		0.48%	0.06%	0.51%	0.21%	4.77%	4.86%	4.87%
FIN	5.24%	6.00%	4.47%	1.19%	1.86%	1.22%	0.48%		0.48%	1.11%	0.77%	4.57%	4.66%	4.67%
GBR	5.08%	5.81%	4.26%	0.99%	1.81%	0.93%	0.06%	0.48%		0.38%	0.27%	4.73%	4.82%	4.84%
IBS	4.69%	5.66%	2.82%	0.42%	1.69%	0.25%	0.51%	1.11%	0.38%		0.40%	7.53%	7.85%	7.71%
TSI	4.98%	5.69%	4.22%	0.98%	1.83%	0.85%	0.21%	0.77%	0.27%	0.40%		4.53%	4.62%	4.64%
CHB	5.78%	6.55%	5.46%	3.51%	3.10%	3.95%	4.77%	4.57%	4.73%	7.53%	4.53%		0.12%	0.45%
CHS	5.85%	6.64%	5.56%	3.63%	3.24%	4.06%	4.86%	4.66%	4.82%	7.85%	4.62%	0.12%		0.60%
JPT	5.78%	6.55%	5.41%	3.54%	3.15%	3.98%	4.87%	4.67%	4.84%	7.71%	4.64%	0.45%	0.60%	

Weir and Cockerham Estimator (MAF > 5%)

	LWK	YRI	ASW	CLM	MXL	PUR	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT
LWK		0.75%	1.25%	9.60%	10.86%	8.90%	11.48%	11.61%	11.48%	11.76%	11.19%	13.36%	13.56%	13.45%
YRI	0.75%		1.31%	10.42%	11.65%	9.71%	12.33%	12.46%	12.34%	12.99%	12.04%	14.12%	14.36%	14.23%
ASW	1.25%	1.31%		5.90%	7.35%	5.18%	8.02%	8.26%	8.05%	6.84%	7.91%	11.22%	11.48%	11.26%
CLM	9.60%	10.42%	5.90%		0.92%	0.52%	1.58%	1.81%	1.58%	1.07%	1.59%	6.81%	7.09%	6.88%
MXL	10.86%	11.65%	7.35%	0.92%		1.79%	3.23%	3.13%	3.23%	2.91%	3.35%	5.75%	6.05%	5.82%
PUR	8.90%	9.71%	5.18%	0.52%	1.79%		1.23%	1.65%	1.26%	0.75%	1.15%	7.66%	7.93%	7.73%
CEU	11.48%	12.33%	8.02%	1.58%	3.23%	1.23%		0.63%	0.06%	0.42%	0.32%	8.82%	9.07%	8.96%
FIN	11.61%	12.46%	8.26%	1.81%	3.13%	1.65%	0.63%		0.64%	1.08%	1.13%	8.14%	8.39%	8.27%
GBR	11.48%	12.34%	8.05%	1.58%	3.23%	1.26%	0.06%	0.64%		0.33%	0.39%	8.81%	9.06%	8.96%
IBS	11.76%	12.99%	6.84%	1.07%	2.91%	0.75%	0.42%	1.08%	0.33%		0.35%	10.74%	11.26%	11.01%
TSI	11.19%	12.04%	7.91%	1.59%	3.35%	1.15%	0.32%	1.13%	0.39%	0.35%		8.71%	8.96%	8.87%
CHB	13.36%	14.12%	11.22%	6.81%	5.75%	7.66%	8.82%	8.14%	8.81%	10.74%	8.71%		0.16%	0.62%
CHS	13.56%	14.36%	11.48%	7.09%	6.05%	7.93%	9.07%	8.39%	9.06%	11.26%	8.96%	0.16%		0.82%
JPT	13.45%	14.23%	11.26%	6.88%	5.82%	7.73%	8.96%	8.27%	8.96%	11.01%	8.87%	0.62%	0.82%	

HapMap Estimator

	LWK	YRI	ASW	CLM	MXL	PUR	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT
LWK		0.37%	0.37%	2.84%	3.88%	2.19%	5.85%	6.61%	6.19%	0.13%	6.68%	8.18%	8.54%	7.48%
YRI	0.37%		0.46%	3.95%	5.12%	3.20%	7.20%	7.92%	7.51%	0.58%	7.93%	9.56%	9.92%	8.91%
ASW	0.37%	0.46%		3.25%	4.73%	2.41%	6.59%	7.12%	6.78%	-0.86%	7.06%	9.66%	9.93%	9.21%
CLM	2.84%	3.95%	3.25%		0.63%	0.34%	1.33%	1.55%	1.34%	-0.33%	1.41%	5.06%	5.25%	4.93%
MXL	3.88%	5.12%	4.73%	0.63%		1.31%	2.02%	1.97%	1.98%	0.94%	2.03%	4.05%	4.25%	3.94%
PUR	2.19%	3.20%	2.41%	0.34%	1.31%		1.43%	1.74%	1.44%	-0.77%	1.47%	5.81%	5.99%	5.71%
CEU	5.85%	7.20%	6.59%	1.33%	2.02%	1.43%		0.32%	0.03%	0.27%	0.15%	5.62%	5.81%	5.59%
FIN	6.61%	7.92%	7.12%	1.55%	1.97%	1.74%	0.32%		0.34%	0.38%	0.60%	5.06%	5.26%	4.99%
GBR	6.19%	7.51%	6.78%	1.34%	1.98%	1.44%	0.03%	0.34%		0.19%	0.20%	5.55%	5.75%	5.51%
IBS	0.13%	0.58%	-0.86%	-0.33%	0.94%	-0.77%	0.27%	0.38%	0.19%		0.15%	3.40%	3.43%	3.61%
TSI	6.68%	7.93%	7.06%	1.41%	2.03%	1.47%	0.15%	0.60%	0.20%	0.15%		5.38%	5.59%	5.32%
CHB	8.18%	9.56%	9.66%	5.06%	4.05%	5.81%	5.62%	5.06%	5.55%	3.40%	5.38%		0.09%	0.35%
CHS	8.54%	9.92%	9.93%	5.25%	4.25%	5.99%	5.81%	5.26%	5.75%	3.43%	5.59%	0.09%		0.47%
JPT	7.48%	8.91%	9.21%	4.93%	3.94%	5.71%	5.59%	4.99%	5.51%	3.61%	5.32%	0.35%	0.47%	

HapMap Estimator (MAF > 5%)

	LWK	YRI	ASW	CLM	MXL	PUR	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT
LWK		0.36%	1.02%	5.84%	6.74%	5.32%	7.48%	7.71%	7.54%	3.00%	7.41%	9.24%	9.44%	9.10%
YRI	0.36%		1.09%	6.70%	7.64%	6.17%	8.31%	8.47%	8.33%	3.77%	8.10%	10.04%	10.24%	9.99%
ASW	1.02%	1.09%		3.64%	4.79%	3.12%	5.41%	5.60%	5.43%	1.77%	5.32%	8.14%	8.34%	8.13%
CLM	5.84%	6.70%	3.64%		0.63%	0.32%	1.16%	1.35%	1.16%	-0.07%	1.20%	4.93%	5.13%	4.92%
MXL	6.74%	7.64%	4.79%	0.63%		1.25%	2.03%	1.96%	2.00%	1.06%	2.03%	4.03%	4.25%	4.01%
PUR	5.32%	6.17%	3.12%	0.32%	1.25%		1.06%	1.34%	1.07%	-0.35%	1.05%	5.53%	5.72%	5.57%
CEU	7.48%	8.31%	5.41%	1.16%	2.03%	1.06%		0.33%	0.03%	0.17%	0.17%	5.87%	6.09%	5.89%
FIN	7.71%	8.47%	5.60%	1.35%	1.96%	1.34%	0.33%		0.34%	0.32%	0.62%	5.29%	5.50%	5.26%
GBR	7.54%	8.33%	5.43%	1.16%	2.00%	1.07%	0.03%	0.34%		0.12%	0.20%	5.82%	6.03%	5.81%
IBS	3.00%	3.77%	1.77%	-0.07%	1.06%	-0.35%	0.17%	0.32%	0.12%		0.08%	3.38%	3.43%	3.67%
TSI	7.41%	8.10%	5.32%	1.20%	2.03%	1.05%	0.17%	0.62%	0.20%	0.08%		5.66%	5.88%	5.63%
CHB	9.24%	10.04%	8.14%	4.93%	4.03%	5.53%	5.87%	5.29%	5.82%	3.38%	5.66%		0.09%	0.34%
CHS	9.44%	10.24%	8.34%	5.13%	4.25%	5.72%	6.09%	5.50%	6.03%	3.43%	5.88%	0.09%		0.46%
JPT	9.10%	9.99%	8.13%	4.92%	4.01%	5.57%	5.89%	5.26%	5.81%	3.67%	5.63%	0.34%	0.46%	

Hudson Definition/Estimator, Ratio of Averages

	LWK	YRI	ASW	CLM	MXL	PUR	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT
LWK		0.79%	1.04%	10.30%	12.02%	9.35%	12.75%	12.92%	12.73%	12.37%	12.36%	15.13%	15.30%	15.24%
YRI	0.79%		0.92%	11.24%	12.97%	10.25%	13.79%	13.97%	13.79%	13.39%	13.41%	16.07%	16.26%	16.19%
ASW	1.04%	0.92%		6.45%	8.10%	5.65%	8.51%	8.71%	8.49%	8.23%	8.25%	11.82%	12.01%	11.94%
CLM	10.30%	11.24%	6.45%		1.05%	0.58%	1.50%	1.77%	1.49%	1.48%	1.50%	7.25%	7.52%	7.36%
MXL	12.02%	12.97%	8.10%	1.05%		2.01%	3.66%	3.57%	3.64%	3.77%	3.76%	6.42%	6.72%	6.49%
PUR	9.35%	10.25%	5.65%	0.58%	2.01%		1.05%	1.51%	1.07%	1.04%	0.95%	8.08%	8.32%	8.21%
CEU	12.75%	13.79%	8.51%	1.50%	3.66%	1.05%		0.65%	0.06%	0.42%	0.36%	10.33%	10.62%	10.51%
FIN	12.92%	13.97%	8.71%	1.77%	3.57%	1.51%	0.65%		0.68%	1.14%	1.20%	9.55%	9.84%	9.71%
GBR	12.73%	13.79%	8.49%	1.49%	3.64%	1.07%	0.06%	0.68%		0.37%	0.42%	10.30%	10.58%	10.48%
IBS	12.37%	13.39%	8.23%	1.48%	3.77%	1.04%	0.42%	1.14%	0.37%		0.40%	10.39%	10.71%	10.55%
TSI	12.36%	13.41%	8.25%	1.50%	3.76%	0.95%	0.36%	1.20%	0.42%	0.40%		10.30%	10.58%	10.49%
CHB	15.13%	16.07%	11.82%	7.25%	6.42%	8.08%	10.33%	9.55%	10.30%	10.39%	10.30%		0.17%	0.67%
CHS	15.30%	16.26%	12.01%	7.52%	6.72%	8.32%	10.62%	9.84%	10.58%	10.71%	10.58%	0.17%		0.90%
JPT	15.24%	16.19%	11.94%	7.36%	6.49%	8.21%	10.51%	9.71%	10.48%	10.55%	10.49%	0.67%	0.90%	

Table S12A Summary of sites showing high levels of population differentiation

LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	317	52.6
CON	ASN-EUR	190	53.4

LEVEL	AFR=Africa; EUR=Europe; ASN=Asia; GLO=global sample
POP_PAIR	Populations pair
CHR	Chromosome
POS	Chromosome position (GRCh37/hg19)
RSID	dbSNP ID (Build 135)
AA	Ancestral allele (uppercase=high confidence; lowercase=low confidence)
DDAF	Derived allele frequencies absolute difference
FUNC_ANN	Functional annotation(s)
HGNC_symbol	HUGO gene name
Ensemlb_GENE_ID	Ensembl gene ID
GENE_START_bp	Gene base start
GENE_END_bp	Gene base end
GENE_PRODUCT	Gene product

* Within Gene_START-Gene_END interval

Table S12B Within-ancestry group high differentiation SNPs (top 10 shown for each comparison)

LEVEL	POP_PAIR	CHR	POS	RSID	AA	DDAF	FUNC_ANN	HGNC_symbol	Ensemlb_GENE_ID	GENE_START_bp	GENE_END_bp	GENE_PRODUCT
AFR	ASW-LWK	8	42,253,960	rs7818866	G	0.432	UTR	VDAC3	ENSG00000078668	42,249,142	42,263,415	protein_coding
AFR	ASW-LWK	11	4,054,405	rs6578434	t	0.421	UTR	STIM1	ENSG00000167323	3,875,757	4,114,439	protein_coding
AFR	ASW-LWK	1	188,792,890	rs73068734	T	0.416	-	-	-	-	-	-
AFR	ASW-LWK	19	56,074,189	rs34551970	T	0.399	-	-	-	-	-	-
AFR	ASW-LWK	13	25,919,996	rs9507502	T	0.397	TFPEAK	NUPL1	ENSG00000139496	25,875,662	25,923,938	protein_coding
AFR	ASW-LWK	2	227,973,892	rs73082223	T	0.397	-	COL4A4	ENSG00000081052	227,867,427	228,028,829	protein_coding
AFR	ASW-LWK	22	45,279,529	rs3747226	a	0.395	-	PHF21B	ENSG00000056487	45,277,042	45,405,880	protein_coding
AFR	ASW-LWK	1	110,201,699	rs506008	t	0.394	SYNONYMOUS	GSTM4	ENSG00000168765	110,198,703	110,208,118	protein_coding
AFR	ASW-LWK	7	83,018,986	rs10232760	T	0.386	-	SEMA3E	ENSG00000170381	82,993,222	83,278,479	protein_coding
AFR	ASW-YRI	2	237,044,077	rs7603279	A	0.389	-	-	-	-	-	-
AFR	ASW-YRI	12	22,499,621	rs7960970	C	0.381	TFPEAK	ST8SIA1	ENSG00000111728	22,216,707	22,589,975	protein_coding
AFR	ASW-YRI	20	59,984,746	rs6061352	G	0.376	-	CDH4	ENSG00000179242	59,827,559	60,512,307	protein_coding
AFR	ASW-YRI	2	227,973,892	rs73082223	T	0.371	-	COL4A4	ENSG00000081052	227,867,427	228,028,829	protein_coding
AFR	ASW-YRI	11	23,719,654	rs12274304	G	0.365	-	-	-	-	-	-
AFR	ASW-YRI	5	147,654,463	rs6887885	A	0.364	-	SPINK13	ENSG00000214510	147,647,743	147,665,817	protein_coding
AFR	ASW-YRI	9	125,694,610	rs1868590	C	0.359	TFPEAK	-	-	-	-	-
AFR	ASW-YRI	3	101,814,628	rs6441645	G	0.359	TFMOTIF	-	-	-	-	-
AFR	ASW-YRI	5	175,177,421	rs6556222	A	0.359	-	-	-	-	-	-
AFR	ASW-YRI	10	27,973,632	rs1907373	A	0.357	-	MKX	ENSG00000150051	27,961,803	28,034,989	protein_coding
AFR	LWK-YRI	12	22,499,621	rs7960970	C	0.475	TFPEAK	ST8SIA1	ENSG00000111728	22,216,707	22,589,975	protein_coding
AFR	LWK-YRI	6	39,709,730	rs307491	t	0.444	-	-	-	-	-	-
AFR	LWK-YRI	20	14,163,707	rs62208177	A	0.424	-	MACROD2	ENSG00000172264	13,976,015	16,033,842	protein_coding
AFR	LWK-YRI	16	213,139	rs61420932	G	0.420	PGENE	HBM	ENSG00000206177	203,891	216,767	protein_coding
AFR	LWK-YRI	1	143,470,807	rs7522380	a	0.419	-	-	-	-	-	-
AFR	LWK-YRI	2	8,209,226	rs2058754	C	0.408	-	-	ENSG00000235665	8,062,556	8,418,214	lincRNA
AFR	LWK-YRI	6	113,191,754	rs2086502	A	0.407	-	-	-	-	-	-
AFR	LWK-YRI	11	34,974,109	rs10734430	G	0.397	-	PDHX	ENSG00000110435	34,937,376	35,042,138	protein_coding
AFR	LWK-YRI	20	42,501,897	rs4812748	A	0.388	-	-	-	-	-	-
AFR	LWK-YRI	22	36,663,213	rs58384577	t	0.374	UTR	APOL1	ENSG00000100342	36,649,056	36,663,576	protein_coding
ASN	CHB-CHS	18	32,247,638	rs1240972	G	0.380	-	DTNA	ENSG00000134769	32,073,254	32,471,808	protein_coding
ASN	CHB-CHS	6	39,709,730	rs307491	t	0.348	-	-	-	-	-	-
ASN	CHB-CHS	10	81,512,832	rs3964382	g	0.334	-	-	-	-	-	-
ASN	CHB-CHS	1	43,370,522	rs61777700	t	0.333	TFMOTIF	-	-	-	-	-
ASN	CHB-CHS	11	39,780,072	rs9667766	A	0.333	-	-	-	-	-	-
ASN	CHB-CHS	12	40,177,313	rs4385961	T	0.330	-	C12orf40	ENSG00000180116	40,019,969	40,302,102	protein_coding
ASN	CHB-CHS	6	79,962,805	rs4706087	G	0.329	-	-	-	-	-	-
ASN	CHB-CHS	9	30,999,670	rs10970027	G	0.326	-	-	-	-	-	-
ASN	CHB-CHS	4	48,867,222	rs12645497	t	0.325	-	-	-	-	-	-
ASN	CHB-CHS	6	70,154,873	rs2479987	C	0.323	-	-	-	-	-	-
ASN	CHB-JPT	11	133,531,655	rs11223548	C	0.423	-	-	-	-	-	-
ASN	CHB-JPT	8	143,764,879	rs2976398	G	0.377	TFPEAK	-	-	-	-	-
ASN	CHB-JPT	7	134,452,557	rs77943343	C	0.339	TFMOTIF	CALD1	ENSG00000122786	134,429,003	134,655,479	protein_coding
ASN	CHB-JPT	2	41,366,779	rs77703766	A	0.336	TFMOTIF	-	-	-	-	-
ASN	CHB-JPT	8	106,509,300	rs60855925	g	0.332	UTR	ZFPM2	ENSG00000169946	106,330,920	106,816,760	protein_coding
ASN	CHB-JPT	16	435,529	rs186934484	t	0.330	-	TMEM8A	ENSG00000129925	420,773	437,113	protein_coding
ASN	CHB-JPT	19	49,092,551	rs10401347	A	0.327	ENHANCER	SULT2B1	ENSG00000088002	49,055,429	49,102,683	protein_coding
ASN	CHB-JPT	9	104,385,873	rs10115450	C	0.325	-	GRIN3A	ENSG00000198785	104,331,635	104,500,862	protein_coding
ASN	CHB-JPT	8	500,785	rs12545856	A	0.314	TFPEAK	-	-	-	-	-
ASN	CHB-JPT	3	69,463,899	rs4428188	G	0.314	-	FRMD4B	ENSG00000114541	69,219,141	69,591,734	protein_coding
ASN	CHS-JPT	14	106,205,022	rs12147642	g	0.543	TFPEAK	IGHG1	ENSG00000211896	106,202,680	106,209,408	IG_C_gene
ASN	CHS-JPT	14	106,019,779	rs28771143	g	0.406	-	-	-	-	-	-
ASN	CHS-JPT	4	124,549,928	rs75958653	G	0.404	-	-	-	-	-	-
ASN	CHS-JPT	3	69,419,614	rs34266487	C	0.390	TFPEAK	FRMD4B	ENSG00000114541	69,219,141	69,591,734	protein_coding
ASN	CHS-JPT	2	197,586,085	rs10172319	T	0.388	TFMOTIF	CCDC150	ENSG00000144395	197,504,278	197,628,214	protein_coding
ASN	CHS-JPT	5	88,187,764	rs304142	C	0.387	-	MEF2C	ENSG00000081189	88,013,975	88,199,922	protein_coding
ASN	CHS-JPT	3	83,851,186	rs4380420	C	0.387	-	-	-	-	-	-
ASN	CHS-JPT	11	133,541,783	rs11223554	T	0.384	-	-	-	-	-	-
ASN	CHS-JPT	2	41,366,779	rs77703766	A	0.380	TFMOTIF	-	-	-	-	-
ASN	CHS-JPT	18	76,985,839	rs12605374	G	0.380	-	ATP9B	ENSG00000166377	76,829,394	77,138,278	protein_coding
EUR	CEU-FIN	11	39,781,515	rs9795509	A	0.404	-	-	-	-	-	-
EUR	CEU-FIN	8	7,281,213	rs139624327	T	0.383	-	-	-	-	-	-
EUR	CEU-FIN	1	17,116,355	rs151218067	G	0.378	-	-	-	-	-	-
EUR	CEU-FIN	2	112,190,331	rs149528480	T	0.356	TFMOTIF	-	ENSG00000172965	111,965,353	112,252,677	processed_transcript
EUR	CEU-FIN	14	19,606,909	rs28477704	C	0.355	-	-	ENSG00000258314	19,606,385	19,643,377	lincRNA
EUR	CEU-FIN	17	4,812,470	rs9905341	G	0.347	TFPEAK	-	-	-	-	-
EUR	CEU-FIN	7	153,687,489	rs144996581	A	0.346	-	DPP6	ENSG00000130226	153,584,182	154,685,995	protein_coding
EUR	CEU-FIN	6	166,660,967	rs9356455	c	0.343	-	-	-	-	-	-
EUR	CEU-FIN	15	50,612,659	rs1972701	C	0.338	-	GABPB1	ENSG00000104064	50,569,389	50,647,605	protein_coding
EUR	CEU-FIN	12	8,131,189	rs7300229	C	0.326	-	NECAP1	ENSG00000089818	7,926,148	8,250,367	protein_coding
EUR	CEU-GBR	12	22,499,621	rs7960970	C	0.316	TFPEAK	ST8SIA1	ENSG00000111728	22,216,707	22,589,975	protein_coding
EUR	CEU-GBR	14	19,606,909	rs28477704	C	0.309	-	-	ENSG00000258314	19,606,385	19,643,377	lincRNA
EUR	CEU-GBR	1	17,116,355	rs151218067	G	0.290	-	-	-	-	-	-
EUR	CEU-GBR	1	149,583,516	rs141282873	T	0.283	-	-	ENSG00000232151	149,575,482	149,651,107	processed_transcript
EUR	CEU-GBR	11	39,780,083	rs189303654	G	0.280	-	-	-	-	-	-
EUR	CEU-GBR	9	41,909,186	rs140215685	C	0.279	-	-	-	-	-	-

EUR	CEU-GBR	12	9,648,921	rs11051289	a	0.276	-	-	ENSG00000214776	9,620,148	9,728,864	pseudogene
EUR	CEU-GBR	2	87,929,798	rs56324656	A	0.276	ENHANCER	-	-	-	-	-
EUR	CEU-GBR	6	10,229,201	rs9465613	c	0.275	-	-	-	-	-	-
EUR	CEU-GBR	7	4,441,248	rs10155898	G	0.269	-	-	-	-	-	-
EUR	CEU-TSI	2	136,608,646	rs4988235	G	0.610	TFPEAK	MCM6	ENSG00000076003	136,597,196	136,633,996	protein_coding
EUR	CEU-TSI	2	135,837,906	rs7570971	A	0.597	TFMOTIF	RAB3GAP1	ENSG00000115839	135,809,835	135,933,964	protein_coding
EUR	CEU-TSI	15	28,365,618	rs12913832	A	0.336	-	HERC2	ENSG00000128731	28,356,186	28,567,298	protein_coding
EUR	CEU-TSI	2	137,622,347	rs1649569	C	0.321	-	THSD7B	ENSG00000144229	137,523,115	138,435,287	protein_coding
EUR	CEU-TSI	3	166,540,488	rs6779741	t	0.320	-	-	-	-	-	-
EUR	CEU-TSI	13	98,264,304	rs4349012	T	0.286	-	-	-	-	-	-
EUR	CEU-TSI	16	88,727,519	rs4782395	T	0.284	-	MVD	ENSG00000167508	88,718,343	88,729,569	protein_coding
EUR	CEU-TSI	8	27,418,443	rs2640722	G	0.280	TFMOTIF	GULOP	ENSG00000234770	27,417,791	27,446,590	pseudogene
EUR	CEU-TSI	7	126,107,032	rs7807889	T	0.279	-	GRM8	ENSG00000179603	126,078,652	126,893,348	protein_coding
EUR	CEU-TSI	15	56,879,880	rs12898998	C	0.279	-	-	ENSG00000260392	56,835,150	56,921,790	sense_overlapping
EUR	FIN-GBR	13	103,855,868	rs9518951	C	0.358	-	-	-	-	-	-
EUR	FIN-GBR	12	32,349,938	rs4931618	T	0.352	TFMOTIF	BICD1	ENSG00000151746	32,259,769	32,536,567	protein_coding
EUR	FIN-GBR	20	9,021,020	rs6118441	C	0.351	-	-	-	-	-	-
EUR	FIN-GBR	2	54,738,392	rs17045941	C	0.341	-	SPTBN1	ENSG00000115306	54,683,422	54,896,812	protein_coding
EUR	FIN-GBR	7	76,505,546	rs7789280	A	0.338	-	UPK3B	ENSG00000243566	76,139,745	76,648,340	protein_coding
EUR	FIN-GBR	20	5,493,842	rs6038189	C	0.330	-	-	-	-	-	-
EUR	FIN-GBR	5	79,416,511	rs6867810	G	0.328	-	SERINC5	ENSG00000164300	79,407,050	79,551,898	protein_coding
EUR	FIN-GBR	21	28,721,810	rs7280320	C	0.327	TFMOTIF	-	-	-	-	-
EUR	FIN-GBR	9	803,158	rs10976679	A	0.327	-	-	-	-	-	-
EUR	FIN-GBR	2	68,349,118	rs11126179	T	0.326	-	-	-	-	-	-
EUR	FIN-TSI	2	136,138,627	rs3940549	a	0.505	-	ZRANB3	ENSG00000121988	135,894,486	136,288,806	protein_coding
EUR	FIN-TSI	2	136,608,646	rs4988235	G	0.484	TFPEAK	MCM6	ENSG00000076003	136,597,196	136,633,996	protein_coding
EUR	FIN-TSI	15	28,365,618	rs12913832	A	0.475	-	HERC2	ENSG00000128731	28,356,186	28,567,298	protein_coding
EUR	FIN-TSI	1	17,116,355	rs151218067	G	0.451	-	-	-	-	-	-
EUR	FIN-TSI	10	5,063,728	rs28375324	A	0.423	TFMOTIF	-	-	-	-	-
EUR	FIN-TSI	2	98,557,575	rs142238274	G	0.420	-	TMEM131	ENSG00000075568	98,372,799	98,612,388	protein_coding
EUR	FIN-TSI	2	98,342,323	rs34149969	C	0.416	-	ZAP70	ENSG00000115085	98,330,023	98,356,325	protein_coding
EUR	FIN-TSI	6	86,047,899	rs7764454	T	0.403	-	-	-	-	-	-
EUR	FIN-TSI	20	48,501,606	rs645544	G	0.394	TFPEAK	SLC9A8	ENSG00000197818	48,429,250	48,508,779	protein_coding
EUR	FIN-TSI	8	7,280,445	rs3958991	G	0.394	TFMOTIF	-	-	-	-	-
EUR	GBR-TSI	2	136,608,646	rs4988235	G	0.634	TFPEAK	MCM6	ENSG00000076003	136,597,196	136,633,996	protein_coding
EUR	GBR-TSI	2	135,755,629	rs1530559	G	0.476	-	YSK4	ENSG00000176601	135,722,061	135,805,038	protein_coding
EUR	GBR-TSI	2	136,991,517	rs12986776	C	0.412	-	-	-	-	-	-
EUR	GBR-TSI	15	28,365,618	rs12913832	A	0.397	-	HERC2	ENSG00000128731	28,356,186	28,567,298	protein_coding
EUR	GBR-TSI	11	39,780,083	rs189303654	G	0.395	-	-	-	-	-	-
EUR	GBR-TSI	6	99,536,850	rs6918521	T	0.378	-	-	-	-	-	-
EUR	GBR-TSI	10	5,063,728	rs28375324	A	0.366	TFMOTIF	-	-	-	-	-
EUR	GBR-TSI	1	17,116,355	rs151218067	G	0.364	-	-	-	-	-	-
EUR	GBR-TSI	8	7,764,420	rs142721326	G	0.329	-	-	-	-	-	-
EUR	GBR-TSI	1	167,582,966	rs146150591	a	0.327	-	-	-	-	-	-

Note IBS excluded due to small sample size

Table S12C Between-continental group sites shown high differentiation (top 10 shown for each comparison)

LEVEL	POP_PAIR	CHR	POS	RSID	AA	DDAF	FUNC_ANN	HGNC_symbol	Ensemlb_GENE_ID	GENE_START_bp	GENE_END_bp	GENE_PRODUCT
CON	AFR-ASN	20	53,252,640	rs6014096	A	0.951	-	DOK5	ENSG00000101134	53,092,136	53,267,710	protein_coding
CON	AFR-ASN	2	72,826,665	rs1596930	A	0.944	-	EXOC6B	ENSG00000144036	72,403,113	73,053,177	protein_coding
CON	AFR-ASN	15	55,936,935	rs12903208	G	0.944	-	PRTG	ENSG00000166450	55,903,744	56,035,288	protein_coding
CON	AFR-ASN	1	159,174,683	rs2814778	T	0.943	TFPEAK,TFMOTIF	DARC	ENSG00000213088	159,173,097	159,176,290	protein_coding
CON	AFR-ASN	2	72,501,137	rs2192015	T	0.941	-	EXOC6B	ENSG00000144036	72,403,113	73,053,177	protein_coding
CON	AFR-ASN	22	46,500,164	rs11702897	C	0.931	TFPEAK	-	ENSG00000197182	46,449,749	46,509,808	protein_coding
CON	AFR-ASN	5	119,745,984	rs6862601	C	0.928	-	-	-	-	-	-
CON	AFR-ASN	20	62,175,996	rs10854170	T	0.927	TFPEAK	SRMS	ENSG00000125508	62,172,163	62,178,857	protein_coding
CON	AFR-ASN	6	105,883,147	rs9486092	G	0.926	-	-	-	-	-	-
CON	AFR-ASN	16	87,404,088	rs889603	C	0.922	TFPEAK	FBXO31	ENSG00000103264	87,362,942	87,425,748	protein_coding
CON	AFR-EUR	1	159,174,683	rs2814778	T	0.940	TFPEAK,TFMOTIF	DARC	ENSG00000213088	159,173,097	159,176,290	protein_coding
CON	AFR-EUR	15	48,392,165	rs1834640	G	0.920	-	-	-	-	-	-
CON	AFR-EUR	5	33,951,693	rs16891982	C	0.919	NON_SYNONYMOUS	SLC45A2	ENSG00000164175	33,944,721	33,984,835	protein_coding
CON	AFR-EUR	1	116,935,068	rs10924081	G	0.901	TFPEAK	ATP1A1	ENSG00000163399	116,915,290	116,952,883	protein_coding
CON	AFR-EUR	4	3,666,494	rs58827274	C	0.896	-	-	-	-	-	-
CON	AFR-EUR	8	145,639,681	rs1871534	G	0.884	NON_SYNONYMOUS	SLC39A4	ENSG00000147804	145,635,126	145,642,279	protein_coding
CON	AFR-EUR	11	19,620,227	rs11025189	C	0.880	TFPEAK	NAV2	ENSG00000166833	19,372,271	20,143,144	protein_coding
CON	AFR-EUR	15	54,976,332	rs2414360	G	0.867	-	-	-	-	-	-
CON	AFR-EUR	17	58,610,478	rs1197095	C	0.867	-	-	ENSG00000259349	58,603,654	58,628,159	antisense
CON	AFR-EUR	9	4,859,106	rs172447	T	0.864	-	RCL1	ENSG00000120158	4,792,869	4,861,064	protein_coding
CON	ASN-EUR	15	48,426,484	rs1426654	G	0.982	NON_SYNONYMOUS	SLC24A5	ENSG00000188467	48,413,169	48,434,869	protein_coding
CON	ASN-EUR	5	33,951,693	rs16891982	C	0.963	NON_SYNONYMOUS	SLC45A2	ENSG00000164175	33,944,721	33,984,835	protein_coding
CON	ASN-EUR	6	2,745,352	rs6927195	G	0.926	TFPEAK	MYLK4	ENSG00000145949	2,663,863	2,751,200	protein_coding
CON	ASN-EUR	2	109,543,883	rs922452	C	0.902	-	EDAR	ENSG00000135960	109,510,927	109,605,828	protein_coding
CON	ASN-EUR	20	568,696	rs6053171	G	0.890	-	-	-	-	-	-
CON	ASN-EUR	3	108,192,751	rs4365635	T	0.846	-	MYH15	ENSG00000144821	108,099,216	108,248,169	protein_coding
CON	ASN-EUR	10	78,894,351	rs2574799	T	0.846	-	KCNMA1	ENSG00000156113	78,637,355	79,398,353	protein_coding
CON	ASN-EUR	2	26,113,913	rs78404020	A	0.844	-	-	-	-	-	-
CON	ASN-EUR	15	28,187,772	rs1545397	A	0.843	-	OCA2	ENSG00000104044	28,000,021	28,344,504	protein_coding
CON	ASN-EUR	17	4,400,392	rs11657785	C	0.842	TFMOTIF	-	-	-	-	-

A full list of sites can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/highly_differentiated_sites/

Table S13   Conservation and polymorphism in KEGG pathways

KEGG Category	Number of Genes	GERP / bp	% SNPs with MAF < 0.5%	% Syn SNPs with MAF < 0.5%	% NonSyn SNPs with MAF < 0.5%	SNPs / kb	LOF / kb	Excess NonSyn / kb with MAF < 0.5%
Graft-versus-host disease	41	-0.0201	52.41%	50.00%	53.65%	6.8976	0.2540	0.3322
Asthma	30	0.0270	53.04%	52.58%	53.27%	11.1484	0.5273	0.1089
Metabolism of xenobiotics by cytochrome P450	69	0.0336	66.65%	61.00%	69.56%	6.8326	0.2911	0.9890
Ribosome	87	0.0872	72.67%	71.89%	73.62%	2.2008	0.0512	0.0617
Drug metabolism - cytochrome P450	71	0.1228	66.35%	63.46%	67.85%	7.5365	0.3359	0.5966
Steroid hormone biosynthesis	54	0.1850	65.88%	57.86%	70.37%	7.0934	0.2496	1.3494
Glycosphingolipid biosynthesis - globo series	14	0.1922	68.28%	62.26%	72.22%	4.0304	0.2707	0.6429
Linoleic acid metabolism	29	0.2121	71.48%	67.02%	73.96%	7.2881	0.3233	0.9860
Allograft rejection	37	0.2504	54.43%	54.62%	54.30%	7.1729	0.1881	-0.0309
Autoimmune thyroid disease	52	0.2705	64.15%	60.45%	66.13%	10.1083	0.2240	0.9452
Other glycan degradation	16	0.3221	72.13%	71.89%	72.29%	7.3826	0.2213	0.0642
alpha-Linolenic acid metabolism	19	0.3371	76.77%	71.59%	80.23%	5.3494	0.2437	0.9743
Primary immunodeficiency	35	0.3829	71.20%	67.66%	73.95%	6.4307	0.0926	0.7039
Intestinal immune network for IgA production	48	0.3878	65.57%	67.06%	64.57%	5.7722	0.1543	-0.2616
Arachidonic acid metabolism	53	0.4022	72.58%	67.10%	75.89%	7.7552	0.2632	1.2917
Glutathione metabolism	44	0.4045	68.09%	64.33%	71.02%	4.9126	0.1181	0.5180
Hematopoietic cell lineage	87	0.4295	71.98%	68.98%	73.95%	6.8845	0.1866	0.6658
Drug metabolism - other enzymes	51	0.4304	66.96%	63.89%	68.74%	8.1855	0.3098	0.6956
Retinol metabolism	64	0.4461	67.51%	65.28%	68.72%	8.2191	0.3381	0.5277
Complement and coagulation cascades	68	0.4541	73.16%	68.10%	76.28%	8.4874	0.1874	1.3453
Ascorbate and aldarate metabolism	25	0.4616	64.75%	58.62%	68.20%	8.7159	0.2977	1.2914
Folate biosynthesis	11	0.4975	63.24%	56.79%	68.27%	5.5394	0.0599	0.8273
Tyrosine metabolism	41	0.5052	74.67%	69.71%	77.85%	6.6317	0.1431	1.0860
Glycosaminoglycan degradation	21	0.5269	70.86%	70.43%	71.16%	5.6035	0.2245	0.0799
One carbon pool by folate	17	0.5403	75.84%	76.64%	75.31%	5.3992	0.0803	-0.1847
Oxidative phosphorylation	129	0.5446	74.37%	70.30%	77.03%	4.3440	0.1240	0.5949
Steroid biosynthesis	17	0.5475	73.40%	67.57%	79.06%	5.7648	0.1533	1.0375
Cytosolic DNA-sensing pathway	56	0.5589	73.08%	70.74%	74.64%	6.4460	0.1976	0.5158
Histidine metabolism	28	0.5673	73.15%	68.80%	75.72%	6.3994	0.2059	0.8921
Cytokine-cytokine receptor interaction	265	0.5791	73.33%	70.38%	75.42%	5.7901	0.1324	0.5767
Glycerolipid metabolism	49	0.5870	74.18%	67.70%	79.11%	4.6746	0.1179	0.9373
Glycosphingolipid biosynthesis - ganglio series	15	0.5895	71.65%	64.81%	76.47%	3.2585	0.0999	0.6328
Butanoate metabolism	34	0.5936	72.31%	70.72%	73.27%	5.7034	0.1692	0.3084
Biosynthesis of unsaturated fatty acids	22	0.5984	75.83%	73.68%	77.48%	3.6264	0.0738	0.2953
Pentose and glucuronate interconversions	28	0.6001	64.25%	57.08%	68.61%	8.0256	0.3033	1.3406
Parkinson's disease	126	0.6272	75.26%	70.99%	78.07%	4.5324	0.1070	0.6659
Glycosphingolipid biosynthesis - lacto and neolacto series	26	0.6298	75.35%	70.44%	78.67%	4.6232	0.1287	0.7672
Primary bile acid biosynthesis	16	0.6348	71.71%	70.99%	72.12%	5.8184	0.1467	0.1437
Fatty acid metabolism	42	0.6457	76.09%	71.36%	78.98%	6.2898	0.1983	1.0393
Sulfur metabolism	13	0.6577	69.23%	68.97%	69.39%	4.9102	0.1469	0.0420
Porphyrin and chlorophyll metabolism	41	0.6585	68.82%	63.87%	71.82%	7.6972	0.1709	1.0545
Nicotinate and nicotinamide metabolism	24	0.6835	75.98%	70.93%	79.40%	4.6301	0.2060	0.8048
PPAR signaling pathway	69	0.6868	77.43%	72.36%	81.00%	5.7948	0.1937	1.0637
Tryptophan metabolism	40	0.6961	74.86%	72.10%	76.70%	7.0455	0.2072	0.6955
Amino sugar and nucleotide sugar metabolism	44	0.6964	74.83%	70.84%	78.03%	4.4905	0.1165	0.6144
Antigen processing and presentation	85	0.6991	62.13%	62.47%	61.89%	5.8430	0.1661	-0.0526
Glycerophospholipid metabolism	77	0.6995	74.50%	69.78%	78.07%	4.3266	0.0984	0.6760
Lysosome	121	0.7024	73.87%	71.41%	75.66%	5.6914	0.1150	0.4910
NOD-like receptor signaling pathway	62	0.7255	75.66%	71.24%	78.87%	5.1817	0.1153	0.7965
Fructose and mannose metabolism	34	0.7261	73.80%	67.26%	79.72%	5.8778	0.1315	1.1738
Phenylalanine metabolism	17	0.7403	70.47%	67.22%	72.80%	8.1204	0.0755	0.8034
RIG-I-like receptor signaling pathway	71	0.7558	74.13%	73.11%	74.83%	4.6367	0.1284	0.1752
Glycolysis / Gluconeogenesis	61	0.7708	75.39%	70.27%	79.19%	6.7050	0.1603	1.1562
Type I diabetes mellitus	43	0.7716	63.55%	58.60%	67.11%	6.6060	0.1288	0.7888
Pyrimidine metabolism	95	0.7720	74.81%	70.47%	78.50%	4.8542	0.1140	0.7146
Valine, leucine and isoleucine degradation	44	0.7738	76.05%	71.71%	78.90%	5.1769	0.1626	0.7948
Pentose phosphate pathway	27	0.7792	75.39%	66.97%	82.97%	7.3695	0.1887	1.8794
SNARE interactions in vesicular transport	37	0.7861	75.66%	74.14%	76.60%	3.0855	0.0812	0.1813
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	25	0.7917	73.86%	69.67%	76.79%	5.5284	0.1305	0.7637
Riboflavin metabolism	16	0.7967	71.85%	62.16%	77.43%	6.6554	0.1808	1.7043
beta-Alanine metabolism	22	0.8052	75.70%	70.89%	78.48%	6.1797	0.1626	1.0211
Terpenoid backbone biosynthesis	15	0.8059	73.73%	68.38%	78.26%	4.2603	0.1002	0.7207
Regulation of autophagy	34	0.8105	72.47%	67.00%	76.19%	4.2222	0.0769	0.6998
Pantothenate and CoA biosynthesis	16	0.8115	74.34%	71.08%	76.22%	5.2944	0.2577	0.5954
O-Glycan biosynthesis	30	0.8121	75.40%	70.83%	78.67%	4.7068	0.1259	0.7371
Pyruvate metabolism	40	0.8181	75.02%	70.06%	78.89%	6.1095	0.1254	1.0124
Apoptosis	87	0.8227	75.16%	71.65%	77.79%	4.5141	0.0748	0.5578
Toll-like receptor signaling pathway	102	0.8251	73.02%	69.32%	75.99%	4.8741	0.1043	0.5885
Leishmania Infection	72	0.8435	70.24%	66.56%	73.44%	5.2868	0.1259	0.5815
Base excision repair	33	0.8530	73.31%	68.70%	76.12%	7.7649	0.2543	1.1442
Sphingolipid metabolism	39	0.8560	72.85%	69.19%	75.69%	5.6615	0.1187	0.6718
Limonene and pinene degradation	10	0.8635	77.27%	78.26%	76.74%	6.4597	0.1468	-0.2936
Peroxisome	78	0.8692	76.86%	72.66%	79.64%	6.0333	0.1452	0.9265
Propanoate metabolism	32	0.8853	75.48%	71.61%	78.18%	5.5815	0.1076	0.7604
Systemic lupus erythematosus	135	0.8868	68.77%	65.31%	71.83%	5.6810	0.1569	0.5664
Renin-angiotensin system	17	0.8880	76.55%	72.29%	79.52%	7.4264	0.1715	1.1418
Nitrogen metabolism	23	0.8999	73.38%	70.33%	75.74%	4.5559	0.1420	0.4686
Selenoamino acid metabolism	25	0.9014	76.06%	72.77%	78.80%	4.2893	0.0745	0.5191
Glycine, serine and threonine metabolism	31	0.9054	74.05%	67.07%	78.52%	7.9198	0.1603	1.6775
Natural killer cell mediated cytotoxicity	133	0.9093	71.40%	66.72%	75.12%	5.2649	0.1137	0.7390
Glyoxylate and dicarboxylate metabolism	16	0.9284	74.74%	71.95%	76.75%	5.6202	0.1147	0.5598
Ether lipid metabolism	33	0.9306	75.10%	66.67%	80.38%	5.0831	0.1599	1.2858
Homologous recombination	28	0.9369	76.78%	71.65%	79.72%	6.5608	0.0802	1.1874
p53 signaling pathway	68	0.9386	73.23%	69.06%	76.50%	3.9781	0.0669	0.5363
Nucleotide excision repair	44	0.9485	78.96%	72.61%	83.46%	6.0749	0.1176	1.4084
Jak-STAT signaling pathway	155	0.9627	75.28%	73.44%	76.72%	5.2934	0.0948	0.3669
Cysteine and methionine metabolism	34	0.9661	73.67%	68.59%	78.24%	4.9437	0.1079	0.7998
Purine metabolism	158	0.9758	75.15%	70.39%	78.99%	5.2612	0.1086	0.8452
Starch and sucrose metabolism	52	0.9784	69.77%	67.25%	71.47%	7.4653	0.2496	0.5750
RNA polymerase	29	0.9798	73.28%	71.93%	74.80%	4.7201	0.1138	0.2277
DNA replication	35	0.9828	79.21%	72.18%	84.19%	7.0908	0.1106	1.7917
Glycosaminoglycan biosynthesis - keratan sulfate	15	0.9845	77.02%	75.81%	78.38%	3.4780	0.0740	0.1746

Glycosaminoglycan biosynthesis - chondroitin sulfate	22	0.9866	69.11%	65.15%	72.36%	5.4593	0.0652	0.6205
Adipocytokine signaling pathway	67	1.0085	77.17%	71.09%	82.63%	4.5105	0.0709	0.9479
Cell adhesion molecules (CAMs)	132	1.0274	72.29%	68.14%	75.79%	6.2160	0.0964	0.8085
Valine, leucine and isoleucine biosynthesis	11	1.0365	78.60%	67.29%	85.39%	6.5327	0.1605	2.2581
Alanine, aspartate and glutamate metabolism	32	1.0465	72.45%	68.45%	75.78%	5.0135	0.0791	0.6360
Glycosaminoglycan biosynthesis - heparan sulfate	26	1.0501	73.30%	65.17%	80.17%	4.0657	0.0385	0.9490
Arginine and proline metabolism	54	1.0515	72.59%	67.90%	76.39%	6.1549	0.1361	0.8992
Vibrio cholerae infection	54	1.0572	73.45%	69.23%	77.39%	5.6519	0.0972	0.7755
Galactose metabolism	26	1.0600	74.30%	71.50%	76.94%	7.2196	0.1750	0.7101
Epithelial cell signaling in Helicobacter pylori infection	68	1.0623	75.32%	71.00%	78.97%	4.3264	0.0946	0.6450
Fc epsilon RI signaling pathway	79	1.0665	73.93%	68.01%	80.18%	3.9882	0.0750	0.7387
Bladder cancer	42	1.0806	75.29%	71.26%	79.53%	4.5657	0.0806	0.6402
N-Glycan biosynthesis	46	1.0841	73.29%	69.37%	76.42%	4.8872	0.0926	0.6245
Neuroactive ligand-receptor interaction	271	1.0877	73.35%	68.18%	77.52%	6.3871	0.1260	1.0363
Olfactory transduction	382	1.0905	61.95%	61.41%	62.18%	16.9448	0.5508	0.2341
Aminoacyl-tRNA biosynthesis	41	1.1020	76.40%	70.91%	80.20%	6.4340	0.0743	1.2150
VEGF signaling pathway	76	1.1127	75.28%	69.48%	80.75%	4.2553	0.0894	0.8086
Taurine and hypotaurine metabolism	10	1.1166	71.19%	64.21%	75.89%	5.5573	0.0235	1.0832
Progesterone-mediated oocyte maturation	85	1.1167	75.50%	68.64%	83.16%	3.9256	0.0534	0.8590
Chemokine signaling pathway	189	1.1259	75.89%	70.01%	81.42%	4.5618	0.0650	0.8956
Alzheimer's disease	163	1.1469	74.75%	69.58%	79.63%	4.8773	0.0858	0.8286
Citrate cycle (TCA cycle)	30	1.1529	71.80%	67.49%	75.40%	7.0320	0.1498	0.9326
Non-small cell lung cancer	54	1.1557	73.87%	69.64%	78.53%	3.4838	0.0296	0.4854
Protein export	23	1.1611	73.48%	71.32%	75.33%	3.1403	0.0338	0.2364
Endocytosis	181	1.1739	74.89%	70.04%	79.24%	5.0552	0.0744	0.8181
mTOR signaling pathway	52	1.1829	76.77%	71.91%	82.90%	3.6746	0.0448	0.6357
Pancreatic cancer	70	1.1879	76.29%	72.61%	79.97%	3.6996	0.0645	0.4974
Leukocyte transendothelial migration	115	1.1906	74.47%	68.91%	79.77%	5.0713	0.0840	0.9073
Circadian rhythm - mammal	13	1.1934	75.83%	72.51%	78.44%	5.9257	0.0247	0.7160
Huntington's disease	177	1.2018	76.07%	73.96%	77.83%	4.9509	0.0855	0.4017
ABC transporters	44	1.2030	75.51%	70.83%	78.52%	9.8768	0.3322	1.5848
Fc gamma R-mediated phagocytosis	96	1.2233	74.30%	69.37%	78.89%	4.7385	0.0727	0.7625
Vascular smooth muscle contraction	115	1.2399	74.67%	70.05%	79.10%	5.0336	0.0831	0.7755
B cell receptor signaling pathway	75	1.2490	73.00%	66.82%	79.25%	4.2512	0.0503	0.7921
Prion diseases	35	1.2599	74.67%	70.71%	78.02%	6.4623	0.1411	0.8734
Mismatch repair	23	1.2617	74.48%	70.30%	76.94%	7.9493	0.2107	1.1178
Chronic myeloid leukemia	73	1.2625	75.41%	71.31%	79.69%	3.5633	0.0467	0.5091
Lysine degradation	44	1.2651	75.30%	72.21%	77.67%	5.7411	0.1106	0.6374
Melanoma	70	1.2677	75.69%	69.53%	82.32%	3.3856	0.0597	0.6848
Proximal tubule bicarbonate reclamation	22	1.2683	74.91%	73.36%	76.51%	6.2701	0.1320	0.3662
Non-homologous end-joining	13	1.2691	78.91%	81.01%	77.55%	5.9237	0.1029	-0.6566
GnRH signaling pathway	101	1.2747	74.35%	68.16%	80.73%	4.8717	0.0805	0.9478
Proteasome	44	1.2819	72.09%	66.29%	78.17%	3.3977	0.0329	0.5851
Dorso-ventral axis formation	24	1.2917	71.91%	65.53%	78.10%	5.8588	0.1105	1.0837
Glioma	65	1.2943	74.06%	68.48%	80.48%	3.6480	0.0459	0.6455
Insulin signaling pathway	137	1.2999	75.55%	71.07%	79.94%	4.8142	0.0677	0.7459
Aldosterone-regulated sodium reabsorption	42	1.3044	73.31%	68.16%	78.98%	4.5925	0.0516	0.7438
T cell receptor signaling pathway	108	1.3076	74.76%	69.27%	80.51%	4.0328	0.0448	0.7198
Neurotrophin signaling pathway	126	1.3087	74.39%	69.10%	79.80%	3.7210	0.0482	0.6367
Thyroid cancer	29	1.3094	76.99%	73.68%	80.50%	4.0868	0.0369	0.5138
Oocyte meiosis	112	1.3109	74.08%	68.84%	79.86%	3.9404	0.0478	0.6622
Small cell lung cancer	84	1.3144	75.63%	70.84%	79.39%	6.8443	0.0692	1.1232
Taste transduction	51	1.3236	72.67%	68.91%	75.71%	8.8760	0.2274	1.0715
Amyotrophic lateral sclerosis (ALS)	51	1.3290	74.44%	67.90%	80.39%	4.4108	0.0858	0.8994
Regulation of actin cytoskeleton	211	1.3379	76.14%	70.87%	81.07%	5.3839	0.0870	0.9740
Maturity onset diabetes of the young	25	1.3509	71.96%	69.23%	74.49%	3.9900	0.1056	0.3536
Phosphatidylinositol signaling system	76	1.3514	74.70%	66.62%	82.36%	5.1850	0.0814	1.2553
Calcium signaling pathway	177	1.3536	74.17%	69.12%	79.34%	6.0295	0.0844	0.9859
Pathways in cancer	324	1.3568	75.73%	71.86%	79.22%	5.4476	0.0689	0.7484
Basal transcription factors	35	1.3641	72.30%	67.72%	76.22%	5.2090	0.1063	0.7394
Pathogenic Escherichia coli infection	54	1.3658	72.41%	69.82%	75.62%	4.6576	0.0797	0.3995
Cell cycle	124	1.3745	76.86%	73.33%	79.85%	4.3522	0.0570	0.5763
Melanogenesis	101	1.3761	77.29%	73.17%	81.73%	4.2687	0.0475	0.6545
ECM-receptor interaction	84	1.3930	74.32%	68.19%	78.84%	10.0136	0.1393	1.9306
Acute myeloid leukemia	57	1.4002	76.24%	73.74%	79.20%	3.5782	0.0474	0.3408
Colorectal cancer	62	1.4035	75.32%	69.82%	80.69%	4.0290	0.0574	0.7352
MAPK signaling pathway	266	1.4086	75.41%	70.11%	80.91%	4.4701	0.0571	0.7932
Cardiac muscle contraction	77	1.4167	73.26%	69.66%	77.51%	5.2025	0.0938	0.6178
ErbB signaling pathway	87	1.4172	74.54%	71.40%	78.06%	3.9630	0.0674	0.4346
Tight junction	131	1.4200	75.50%	69.86%	80.45%	5.9718	0.0928	1.1170
Endometrial cancer	52	1.4222	74.36%	68.26%	80.62%	4.2673	0.0503	0.8197
Prostate cancer	88	1.4256	77.45%	73.46%	81.73%	4.1382	0.0527	0.6226
Notch signaling pathway	47	1.4403	73.23%	67.93%	78.40%	6.0744	0.0551	1.0041
Gap junction	89	1.4459	74.99%	68.81%	82.40%	4.9384	0.0579	0.9779
Focal adhesion	199	1.4481	74.99%	69.60%	79.73%	7.0446	0.0945	1.2483
RNA degradation	56	1.4499	75.70%	73.06%	77.91%	3.7303	0.0556	0.3652
Inositol phosphate metabolism	54	1.4548	75.45%	67.00%	82.65%	5.7267	0.1008	1.4663
Long-term depression	70	1.4718	74.67%	68.19%	82.33%	5.1857	0.0923	1.0563
Renal cell carcinoma	70	1.4865	76.25%	73.49%	79.66%	3.5419	0.0369	0.3685
Type II diabetes mellitus	47	1.4881	73.88%	68.74%	80.08%	4.9770	0.0408	0.8180
Ubiquitin mediated proteolysis	134	1.4933	75.98%	71.71%	80.48%	4.2878	0.0610	0.6472
Adherens junction	73	1.5145	76.36%	71.45%	81.44%	5.1530	0.0674	0.8853
Viral myocarditis	70	1.5208	73.23%	69.24%	76.50%	9.3327	0.2142	1.2103
Long-term potentiation	70	1.5580	75.30%	70.45%	81.75%	4.0585	0.0552	0.6662
Wnt signaling pathway	150	1.5742	75.54%	71.16%	80.11%	4.0333	0.0402	0.6122
Basal cell carcinoma	55	1.5831	76.21%	72.84%	79.47%	5.7745	0.0471	0.7174
Vasopressin-regulated water reabsorption	44	1.6037	77.58%	74.37%	80.89%	4.8000	0.0615	0.6023
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	1.6068	75.98%	69.94%	81.28%	6.6280	0.1140	1.3319
Axon guidance	129	1.6074	74.87%	69.18%	80.24%	5.3155	0.0620	0.9819
Hedgehog signaling pathway	56	1.6267	74.05%	70.55%	77.15%	5.8334	0.0637	0.6942
Spliceosome	124	1.6316	72.76%	69.76%	77.38%	3.5500	0.0665	0.3523
TGF-beta signaling pathway	85	1.6711	76.47%	73.67%	79.10%	4.2332	0.0410	0.4502
Dilated cardiomyopathy	90	1.7720	75.36%	69.47%	80.14%	6.7394	0.0954	1.3019
Hypertrophic cardiomyopathy (HCM)	83	1.8657	74.90%	68.57%	79.79%	6.8656	0.1022	1.3815

Table S14 Average numbers of potentially functional variants per individual in each population

With GERP >2		African			European					Asian			American			Summary	
		ASW	LWK	YRI	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT	CLM	MXL	PUR	Min	Max
synonymous	DAF <0.5%	97	117	103	34	30	33	28	40	41	43	47	42	37	44	28	117
	DAF 0.5-5%	337	399	418	102	103	102	103	104	85	82	86	125	108	139	82	418
	DAF >5%	1312	1274	1255	1393	1394	1390	1390	1394	1401	1399	1400	1405	1402	1394	1255	1405
nonsynonymous	DAF <0.5%	326	404	351	175	162	165	131	193	204	212	221	189	184	202	131	404
	DAF 0.5-5%	747	874	905	318	321	320	318	312	251	236	248	361	320	377	236	905
	DAF >5%	2470	2383	2329	2739	2737	2739	2711	2748	2732	2719	2728	2744	2740	2715	2329	2748
Stop-loss	DAF <0.5%	1.1	1.2	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.2	1.1	1.0	1.0	1.0	1.0	1.2
	DAF 0.5-5%	1.5	1.9	1.8	1.1	1.0	1.2	1.0	1.1	1.0	1.2	1.0	1.1	1.0	1.2	1.0	1.9
	DAF >5%	2.5	2.7	2.7	2.6	2.7	2.6	2.8	2.6	2.2	2.1	2.1	2.3	2.2	2.5	2.1	2.8
HGMD DM	DAF <0.5%	3.8	3.3	3.9	4.8	3.9	4.5	3.4	5.1	2.5	2.5	2.8	3.8	3.1	4.3	2.5	5.1
	DAF 0.5-5%	14	15	17	8.1	8.5	7.6	9.1	7.4	4.8	4.9	4.9	8.5	8.0	8.7	4.8	17
	DAF >5%	16	18	15	11	12	12	12	11	16	16	16	13	13	13	11	18
COSMIC	DAF <0.5%	1.6	2.0	1.7	1.4	1.3	1.5	1.3	1.5	1.4	1.3	1.8	1.5	1.5	1.5	1.3	2.0
	DAF 0.5-5%	4.2	5.1	4.5	1.9	2.1	1.9	1.9	1.8	1.8	2.2	1.9	2.1	1.9	2.2	1.8	5.1
	DAF >5%	9.0	10	9.1	5.6	5.2	5.6	6.1	6.0	6.4	6.7	6.9	5.9	6.2	6.0	5.2	10
UTR	DAF <0.5%	341	122	144	157	158	123	122	121	169	430	140	166	134	367	121	430
	DAF 0.5-5%	1175	403	317	304	484	409	402	417	314	1355	423	527	397	1421	304	1355
	DAF >5%	3701	3968	3937	3927	3973	3977	3973	3956	3924	3530	3950	3951	3971	3492	3530	3977
Non-coding RNA	DAF <0.5%	13	17	14	4.0	4.1	3.9	4.2	4.3	5.1	5.9	6.2	5.4	5.0	5.5	3.9	17.4
	DAF 0.5-5%	58	65	70	18	18	18	17	18	15	14	15	22	19	24	13.7	69.8
	DAF >5%	187	185	179	194	195	192	190	191	191	193	191	196	193	194	178.6	196
Motif_gain_in_TF_peak	DAF <0.5%	11	14	12	5.1	4.7	5.0	5.0	5.4	4.9	4.9	5.8	5.9	5.3	6.6	4.7	14
	DAF 0.5-5%	51	58	59	25	25	25	27	24	23	24	24	28	26	29	23	59
	DAF >5%	174	173	173	171	172	172	168	171	170	169	166	174	167	173	166	174
Motif_loss_in_TF_peak	DAF <0.5%	54	69	59	18	19	18	22	22	21	22	27	24	21	26	18	69
	DAF 0.5-5%	244	281	301	84	84	86	80	81	71	72	73	101	87	109	71	301
	DAF >5%	615	589	584	650	650	650	647	654	637	643	636	649	637	651	584	654
Other conserved	DAF <0.5%	7,641	9,936	8,057	2,026	2,200	2,152	2,510	2,217	2,479	2,821	3,066	3,003	2,600	3,256	2,026	9,936
	DAF 0.5-5%	32,196	37,221	39,359	8,567	8,625	8,620	9,015	8,673	7,162	7,096	7,123	11,048	9,584	12,565	7,096	39,359
	DAF >5%	128,100	124,101	122,904	133,200	133,978	133,431	133,045	133,363	133,341	133,059	132,769	134,155	133,981	133,941	122,904	134,155
Total conserved	DAF <0.5%	8,391	10,871	8,861	2,337	2,500	2,453	2,783	2,564	2,844	3,209	3,483	3,377	2,942	3,652	2,337	10,871
	DAF 0.5-5%	34,543	39,946	42,216	9,399	9,465	9,454	9,862	9,496	7,829	7,739	7,788	12,040	10,453	13,630	7,739	42,216
	DAF >5%	139,994	135,684	134,371	145,666	146,481	145,905	145,475	145,843	145,779	145,468	145,164	146,657	146,442	146,378	134,371	146,657

Without GERP filter		African			European					Asian			American			Summary	
		ASW	LWK	YRI	CEU	FIN	GBR	IBS	TSI	CHB	CHS	JPT	CLM	MXL	PUR	Min	Max
synonymous	DAF <0.5%	506	642	547	175	166	167	139	203	218	226	240	215	194	237	139	642
	DAF 0.5-5%	2005	2359	2468	594	594	591	586	592	497	478	500	740	641	810	478	2468
	DAF >5%	12650	12309	12190	13237	13243	13216	13196	13232	13077	13058	13067	13284	13246	13232	12190	13284
nonsynonymous	DAF <0.5%	645	806	697	298	279	280	224	331	355	365	387	340	320	357	224	806
	DAF 0.5-5%	1954	2278	2377	709	717	711	715	703	563	540	571	833	740	892	540	2377
	DAF >5%	10496	10195	10056	11173	11170	11183	11130	11198	11026	11008	11012	11171	11104	11124	10056	11198
Indel-non-frameshift	DAF <0.5%	1.1	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.1	1.0	1.1	1.0	1.3	1.1	1.0	1.3
	DAF 0.5-5%	19	22	23	6.4	7.1	6.6	6.3	5.6	5.3	5.1	5.1	8.3	7.0	8.4	5.1	23
	DAF >5%	79	73	74	88	88	89	88	87	84	83	84	85	84	86	73	89
Stop-loss	DAF <0.5%	1.5	1.7	1.3	1.1	1.1	1.2	1.3	1.1	1.2	1.3	1.3	1.1	1.1	1.1	1.1	1.7
	DAF 0.5-5%	3.2	3.9	4.3	1.5	1.4	1.4	1.3	1.4	1.1	1.2	1.1	1.5	1.3	1.7	1.1	4.3
	DAF >5%	38	37	37	38	38	39	37	38	38	38	38	39	40	39	37	40
Stop-gain	DAF <0.5%	8.8	10	9.6	5.0	4.7	4.6	3.9	5.5	5.7	5.8	5.5	5.8	4.8	5.8	3.9	10
	DAF 0.5-5%	16	19	18	6.4	6.5	6.3	6.5	6.6	5.8	5.3	5.8	7.5	7.2	7.2	5.3	19
	DAF >5%	25	27	25	26	27	27	24	26	27	27	27	28	28	26	24	28
Indel-frameshift	DAF <0.5%	1.0	1.0	1.0	1.0	1.1	1.1	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.1
	DAF 0.5-5%	15	18	19	5.6	5.8	6.5	7.0	6.0	5.5	5.8	5.8	5.7	5.2	7.3	5.2	19
	DAF >5%	26	25	25	28	28	29	28	29	26	26	26	27	26	29	25	29
Splice site donor	DAF <0.5%	2.9	3.6	3.0	2.1	1.7	1.8	1.7	2.3	2.0	2.4	2.3	1.9	1.9	2.1	1.7	3.6
	DAF 0.5-5%	5.6	7.2	6.6	2.9	2.7	2.8	3.0	2.5	2.4	2.5	2.4	3.5	3.6	3.4	2.4	7.2
	DAF >5%	3.4	3.3	2.6	4.2	4.4	3.9	3.1	4.4	5.1	4.9	5.2	4.3	4.3	4.4	2.6	5.2
Splice site acceptor	DAF <0.5%	2.2	2.9	2.5	1.7	1.7	1.8	1.5	1.9	1.8	1.8	1.6	1.5	1.7	2.0	1.5	2.9
	DAF 0.5-5%	3.6	3.8	4.0	1.7	1.5	1.5	1.8	1.7	1.7	1.5	1.7	1.8	1.6	1.8	1.5	4.0
	DAF >5%	2.2	2.7	2.1	3.3	3.3	3.4	3.4	3.3	4.6	4.3	4.6	3.9	3.6	3.2	2.1	4.6
HGMD-DM	DAF <0.5%	6.2	5.7	6.3	7.2	6.6	7.2	5.0	7.7	3.7	3.8	4.5	5.3	4.6	6.3	3.7	7.7
	DAF 0.5-5%	28	32	33	15	16	16	17	15	11	10	11	16	15	17	10	33
	DAF >5%	39	43	40	28	29	29	30	29	34	34	34	31	30	31	28	43
COSMIC	DAF <0.5%	2.9	3.8	2.9	1.8	1.8	1.7	1.6	1.8	2.0	2.1	3.0	2.2	1.8	2.2	1.6	3.8
	DAF 0.5-5%	12	15	14	4.3	3.8	3.8	4.0	3.8	3.7	3.9	4.2	4.6	3.7	4.9	3.7	15
	DAF >5%	32	32	31	28	28	28	29	28	29	29	30	28	29	29	28	32
UTR	DAF <0.5%	1612	2099	1740	493	498	487	514	559	584	635	715	674	586	714	487	2099
	DAF 0.5-5%	7277	8472	8889	2087	2132	2099	2184	2097	1701	1665	1721	2642	2323	2953	1665	8889
	DAF >5%	42816	41488	41139	44839	44905	44854	44824	44873	44201	44141	44152	44898	44746	44895	41139	44905
Non-coding_RNA	DAF <0.5%	160	210	173	44	46	44	47	51	52	57	65	61	54	64	44	210
	DAF 0.5-5%	770	898	934	207	210	212	225	212	166	159	170	269	236	313	159	934
	DAF >5%	5007	4880	4826	5208	5222	5229	5227	5217	5143	5152	5136	5240	5183	5240	4826	5240
Motif_gain_in_TF_peak	DAF <0.5%	132	148	136	91	91	90	91	93	92	93	98	96	93	99	90	148
	DAF 0.5-5%	592	650	657	382	389	384	388	383	371	371	374	404	394	419	371	657
	DAF >5%	3779	3829	3823	3329	3333	3331	3330	3350	3342	3341	3336	3429	3389	3471	3329	3829
Motif_loss_in_TF_peak	DAF <0.5%	328	421	349	125	127	124	134	142	136	146	162	157	139	163	124	421
	DAF 0.5-5%	1674	1944	2028	539	557	544	555	533	455	459	465	656	596	732	455	2028
	DAF >5%	5160	5042	4989	5179	5189	5173	5165	5199	5120	5116	5111	5234	5197	5265	4989	5234
all_nonfunction_sites	DAF <0.5%	113,190	148,364	119,718	29,023	31,919	30,960	35,173	32,271	34,875	39,689	43,686	43,872	37,491	47,501	29,023	148,364
	DAF 0.5-5%	545,984	634,101	667,345	140,160	142,839	142,951	151,307	142,224	116,542	115,084	116,942	184,452	159,504	209,087	115,084	667,345
	DAF >5%	3,652,026	3,545,318	3,519,625	3,774,613	3,794,886	3,783,796	3,775,778	3,776,575	3,733,106	3,724,728	3,717,790	3,789,739	3,776,944	3,787,751	3,519,625	3,794,886
all_sites	DAF <0.5%	116,216	152,267	122,978	30,034	32,918	31,944	36,129	33,425	36,081	40,974	45,107	45,184	38,654	48,892	30,034	152,267
	DAF 0.5-5%	558,996	649,303	683,239	143,987	146,737	146,802	155,270	146,066	119,668	118,137	120,115	189,257	163,737	214,430	118,137	683,239
	DAF >5%	3,728,104	3,619,039	3,592,643	3,854,212	3,874,781	3,863,507	3,855,295	3,856,263	3,811,719	3,803,199	3,796,180	3,869,561	3,856,439	3,867,461	3,592,643	3,874,781

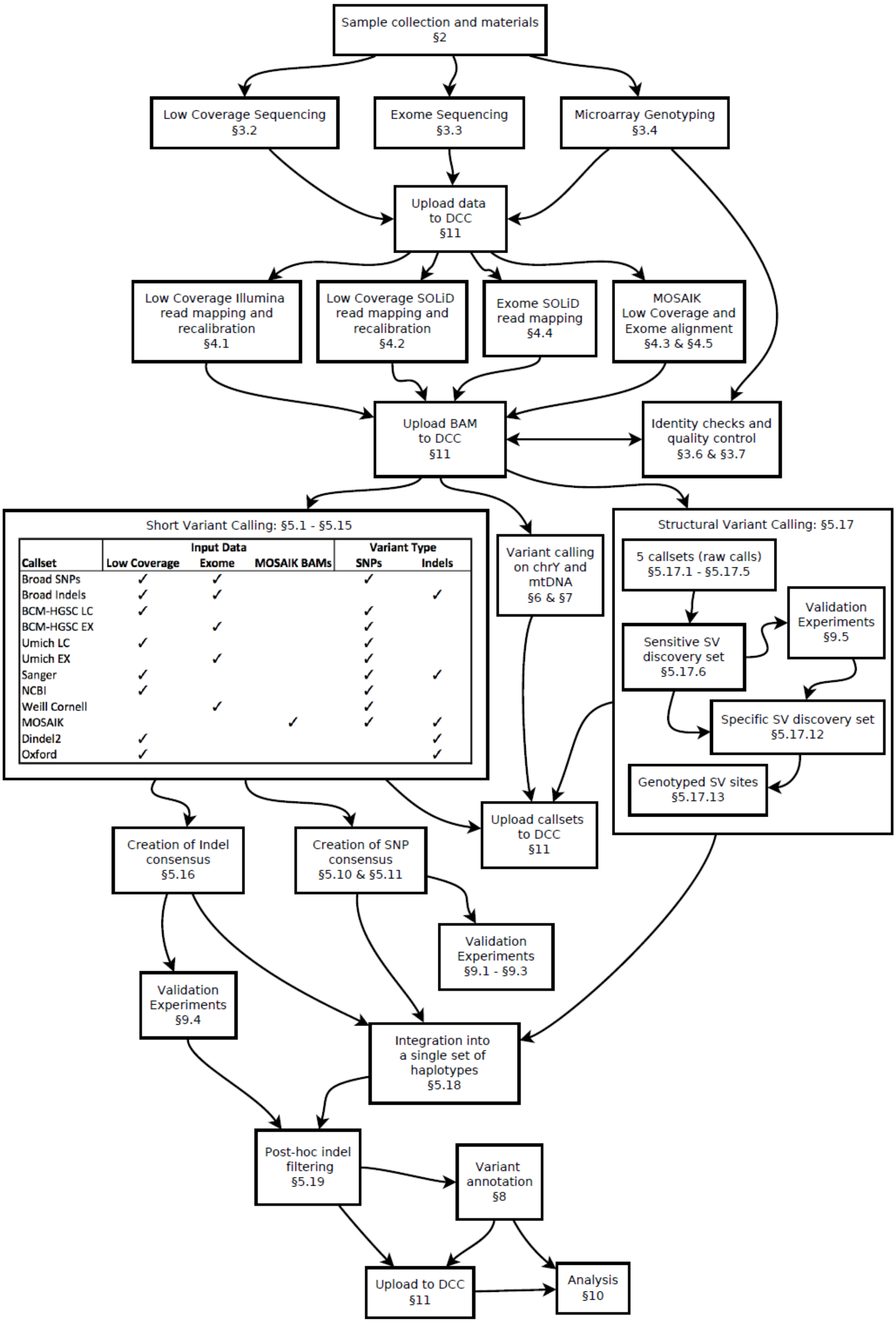
Functional annotations for variants can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/annotated_vcfs/

Conservation scores for variants can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/variant_gerp_scores/

A list of genome annotations used for assigning functional consequence can be found at
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/functional_annotation/annotation_sets/

Table S15 Number of variants in linkage disequilibrium (LD) with the SNPs in GWAS catalog

LD Criteria	Avg. # Variants in LD			%GERP≥2
	HapMap	Pilot	Phase 1	
r^2 ≥ .5 in Africans (n=185)	8.3	18	22.6	4.7
r^2 ≥ .5 in Americans (n=242)	13.3	28.6	35.7	4.7
r^2 ≥ .5 in Asians (n=286)	21.3	46.2	58	4.7
r^2 ≥ .5 in Europeans (n=379)	20.5	44.7	55.8	4.6
r^2 ≥ .5 in all individuals (n=1,092)	14	29.4	36.2	4.8
r^2 ≥ .5 in each continental population	5.9	11.8	14.4	4.7
|D’|=1 in each continental population	10.9	36.3	73	4.7


Sample collection and materials §2
Low Coverage Sequencing §3.2
Exome Sequencing §3.3
Microarray Genotyping §3.4
Upload data to DCC §11
Low Coverage Illumina read mapping and recalibration §4.1
Low Coverage SOLiD read mapping and recalibration §4.2
Exome SOLiD read mapping §4.4
MOSAIK Low Coverage and Exome alignment §4.3 & §4.5
Upload BAM to DCC §11
Identity checks and quality control §3.6 & §3.7
Short Variant Calling: §5.1 - §5.15
Variant calling on chrY and mtDNA §6 & §7
Structural Variant Calling: §5.17
5 callsets (raw calls) §5.17.1 - §5.17.5
Sensitive SV discovery set §5.17.6
Validation Experiments §9.5
Specific SV discovery set §5.17.12
Genotyped SV sites §5.17.13
Upload callsets to DCC §11
Creation of Indel consensus §5.16
Creation of SNP consensus §5.10 & §5.11
Validation Experiments §9.1 - §9.3
Validation Experiments §9.4
Integration into a single set of haplotypes §5.18
Post-hoc indel filtering §5.19
Variant annotation §8
Upload to DCC §11
Analysis §10

	Input Data			Variant Type	
Callset	Low Coverage	Exome	MOSAIK BAMs	SNPs	Indels
Broad SNPs	✓	✓		✓	
Broad Indels	✓	✓			✓
BCM-HGSC LC	✓			✓	
BCM-HGSC EX		✓		✓	
Umich LC	✓			✓	
Umich EX		✓		✓	
Sanger	✓			✓	✓
NCBI	✓			✓	
Weill Cornell		✓		✓	
MOSAIK			✓	✓	✓
Dindel2	✓				✓
Oxford	✓				✓

Figure S1. Overview of data generation, processing and analysis
Flowchart summarising steps involved in generating the 1000 Genomes Project Phase 1 release. Boxes indicate steps in the process and numbers indicate the corresponding section(s) within the supplementary material.

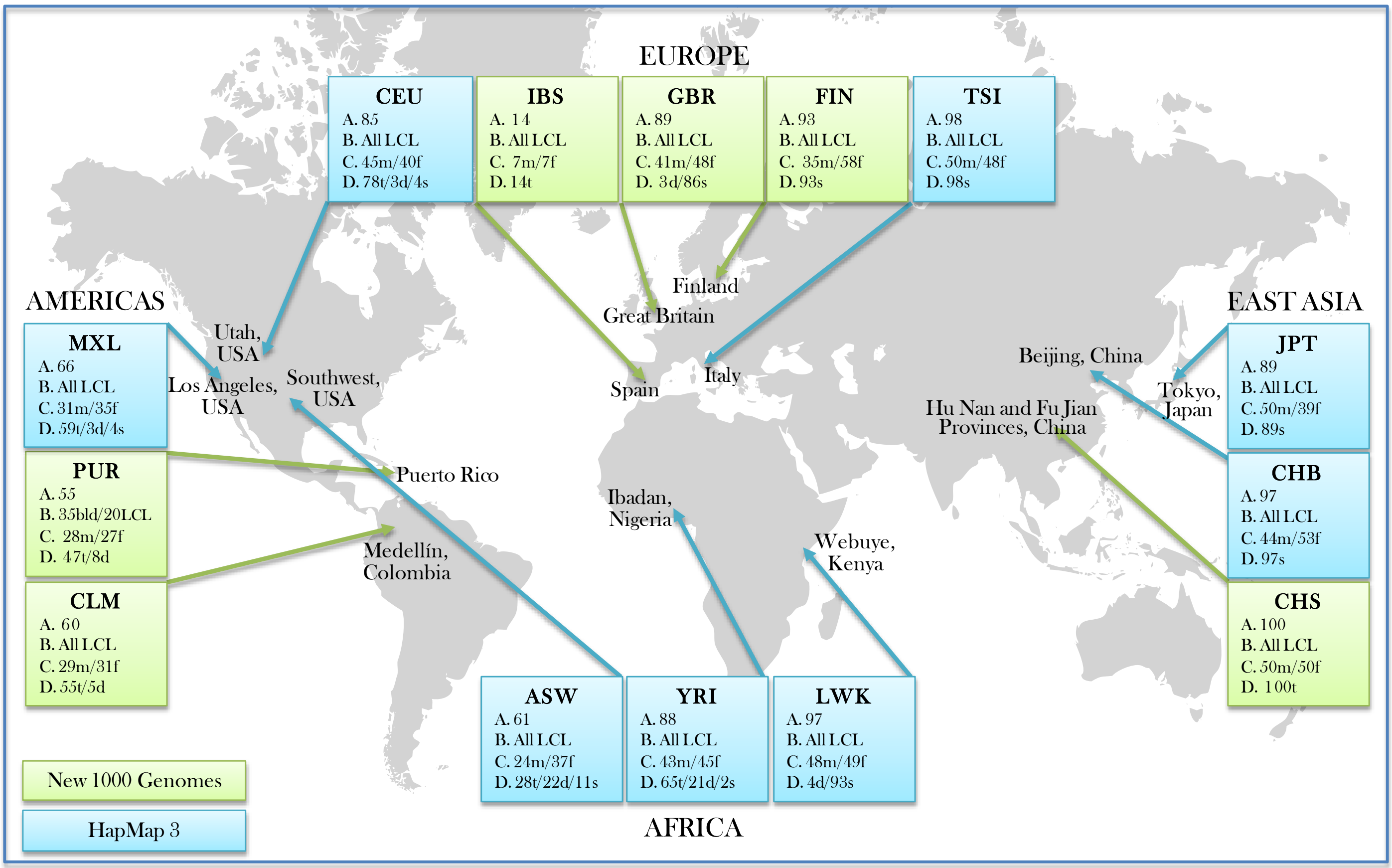


Figure S2. 1000 Genomes Project Phase I populations

Populations collected as part of the HapMap Project (blue) and the 1000 Genomes Project (green) include: Europe (**IBS** (Iberian Populations in Spain), **GBR** (British from England and Scotland), **CEU** (Utah residents with ancestry from northern and western Europe), **FIN** (Finnish in Finland), **TSI** (Toscani in Italia)); East Asia (**JPT** (Japanese in Tokyo, Japan), **CHB** (Han Chinese in Beijing, China), **CHS** (Han Chinese South)); Africa (**ASW** (African Ancestry in SW USA), **YRI** (Yoruba in Ibadan, Nigeria), **LWK** (Luhya in Webuye, Kenya)); Americas (**MXL** (Mexican Ancestry in Los Angeles, CA, USA), **PUR** (Puerto Ricans in Puerto Rico), **CLM** (Colombians in Medellín, Colombia)). **A** – Total number of samples sequenced; **B** – Source of DNA (blood (bld) or LCL); **C** – Gender composition (Male/Female); **D** – Number that are part of mother-father-child trios (t), parent-child duos (d) or singletons (s); for trios and duos, only parent samples were sequenced.

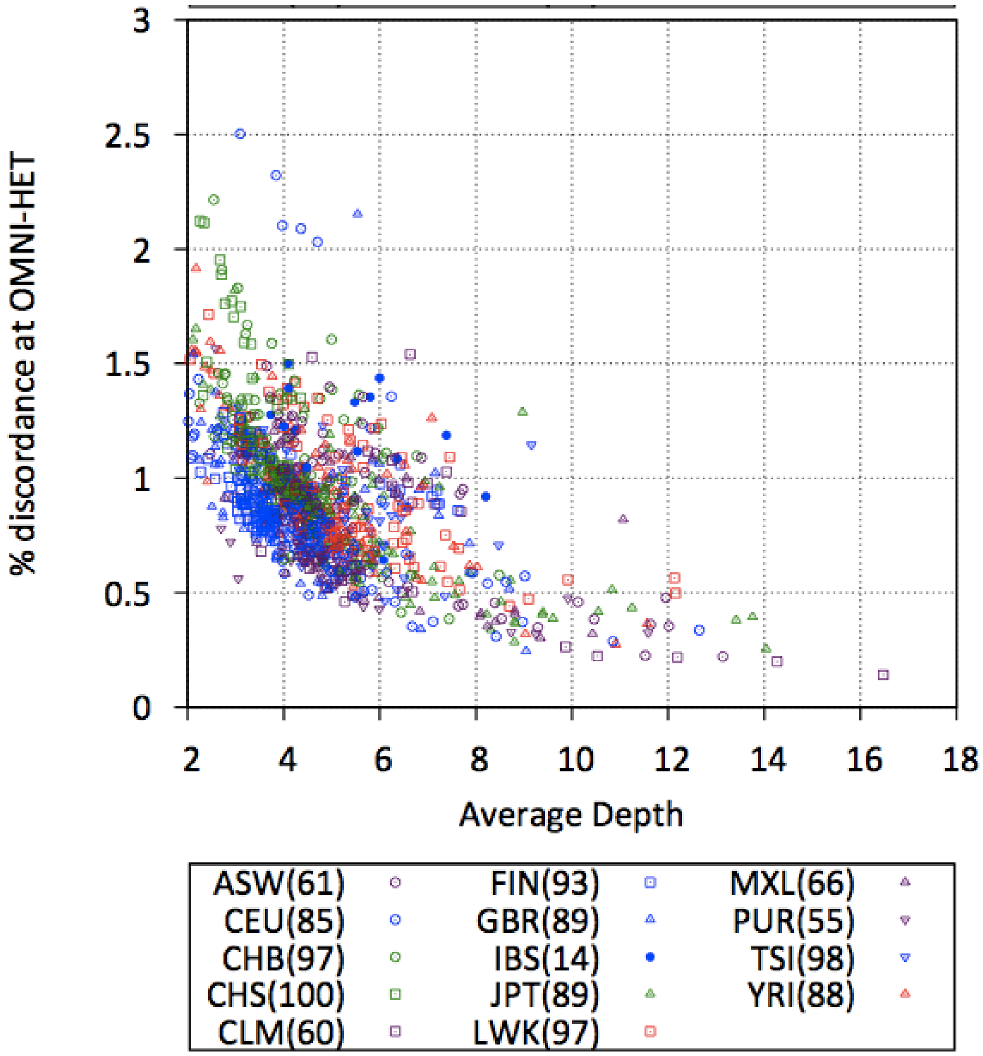


Figure S3. Sequencing depth and genotyping accuracy
The relationship between average sequencing depth (low-coverage data) and genotype discordance between Phase 1 release genotypes and estimates from the OMNI SNP array data at heterozygous sites (identified from the array). Colours indicate the population for each individual.

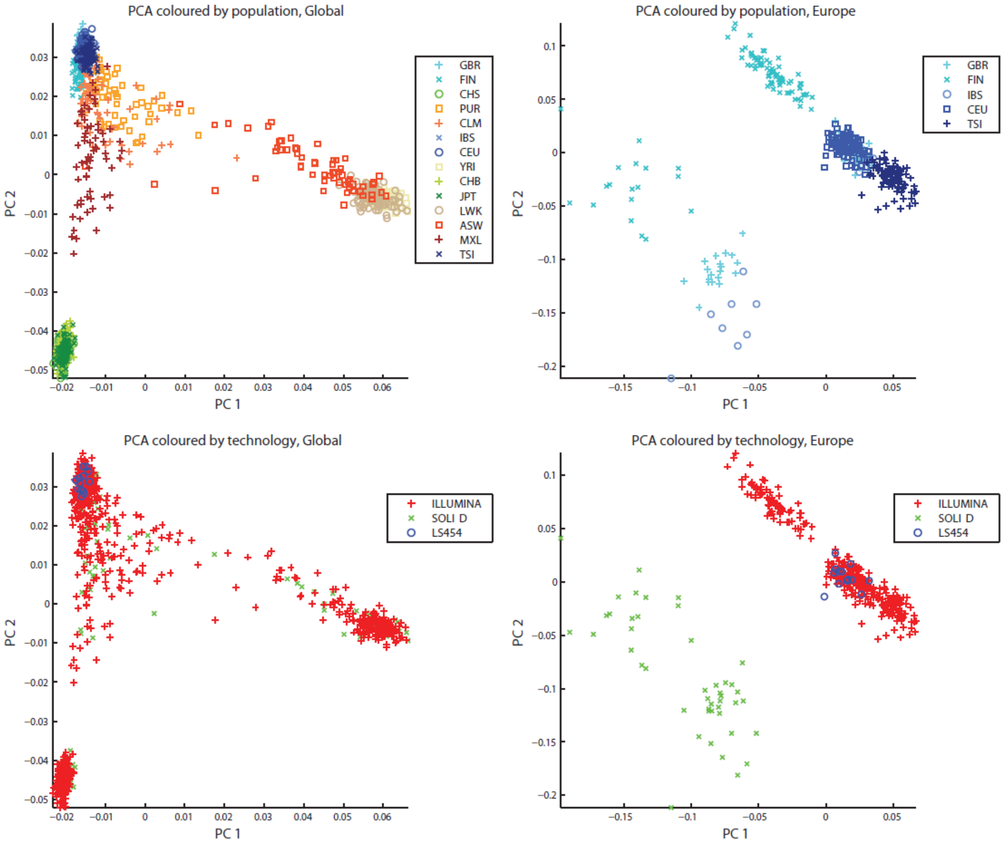


Figure S4. Geography and technology stratify patterns of genetic variation.
PCA plots (1st and 2nd components: estimated from release genotypes, see Methods) for all samples (left hand side) and those within EUR (right hand side). In the top row individuals are coloured by population of origin. In the bottom row samples are coloured by primary technology from which low-coverage data have been generated. At the continental level, the PCA plots mirror previous observations regarding to the relationships between groups. Within Europe, however, technology is an important component driving differentiation between the release haplotypes.

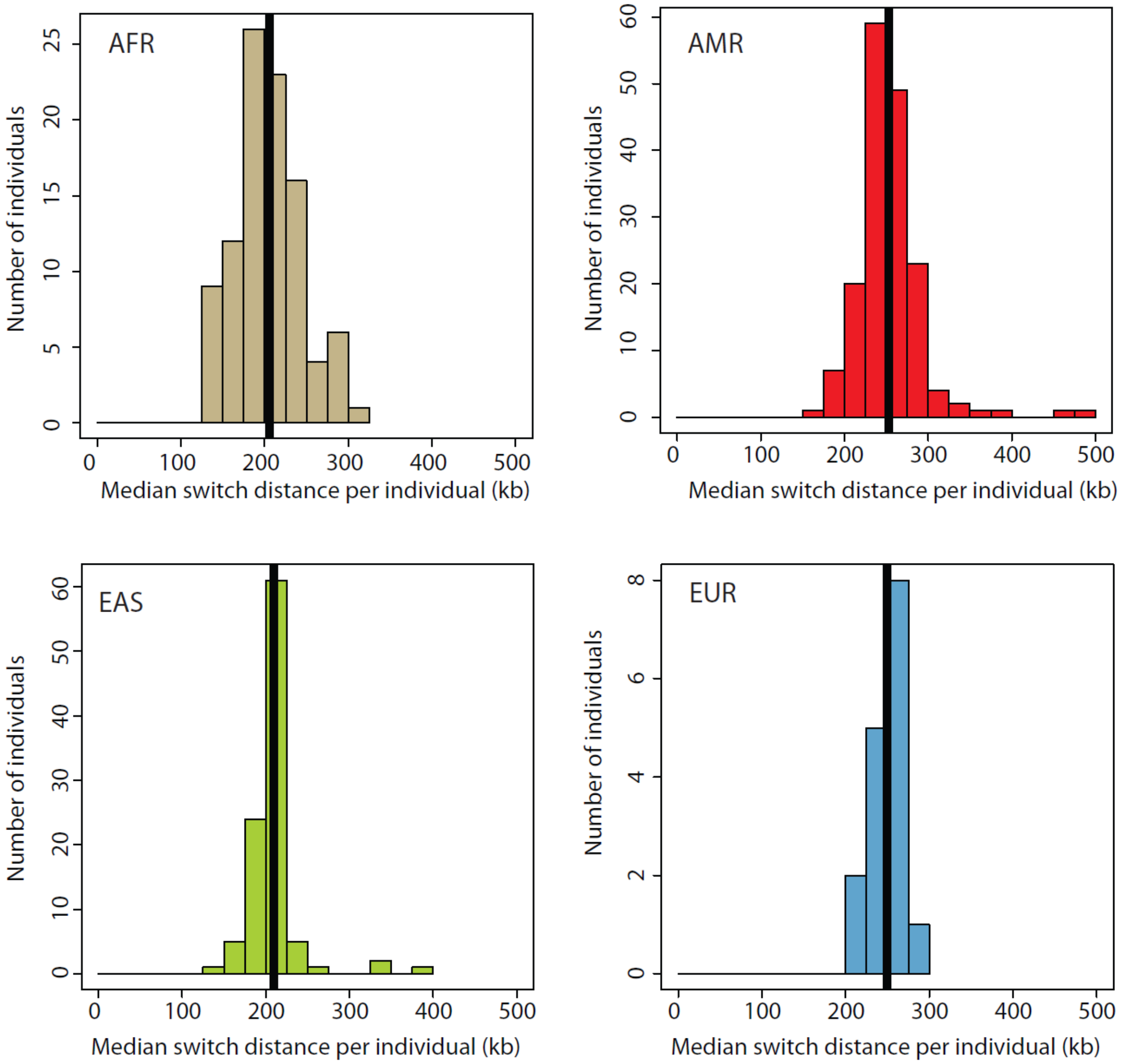


Figure S5. Errors in haplotype estimation.
Distributions of median distance between phase ‘switch errors’ at common SNPs in Phase 1 haplotypes as estimated from comparison to SNP array genotypes (OMNI) genotyped in trios, where haplotypes can be determined by transmission. Trio genotypes were available for 97 individuals from AFR (24 ASW, 73 YRI), 169 individuals from AMR (60 CLM, 54 MXL, 55 PUR), 100 individuals from EAS (100 CHS), and 16 individuals from EUR (2 CEU, 14 IBS).

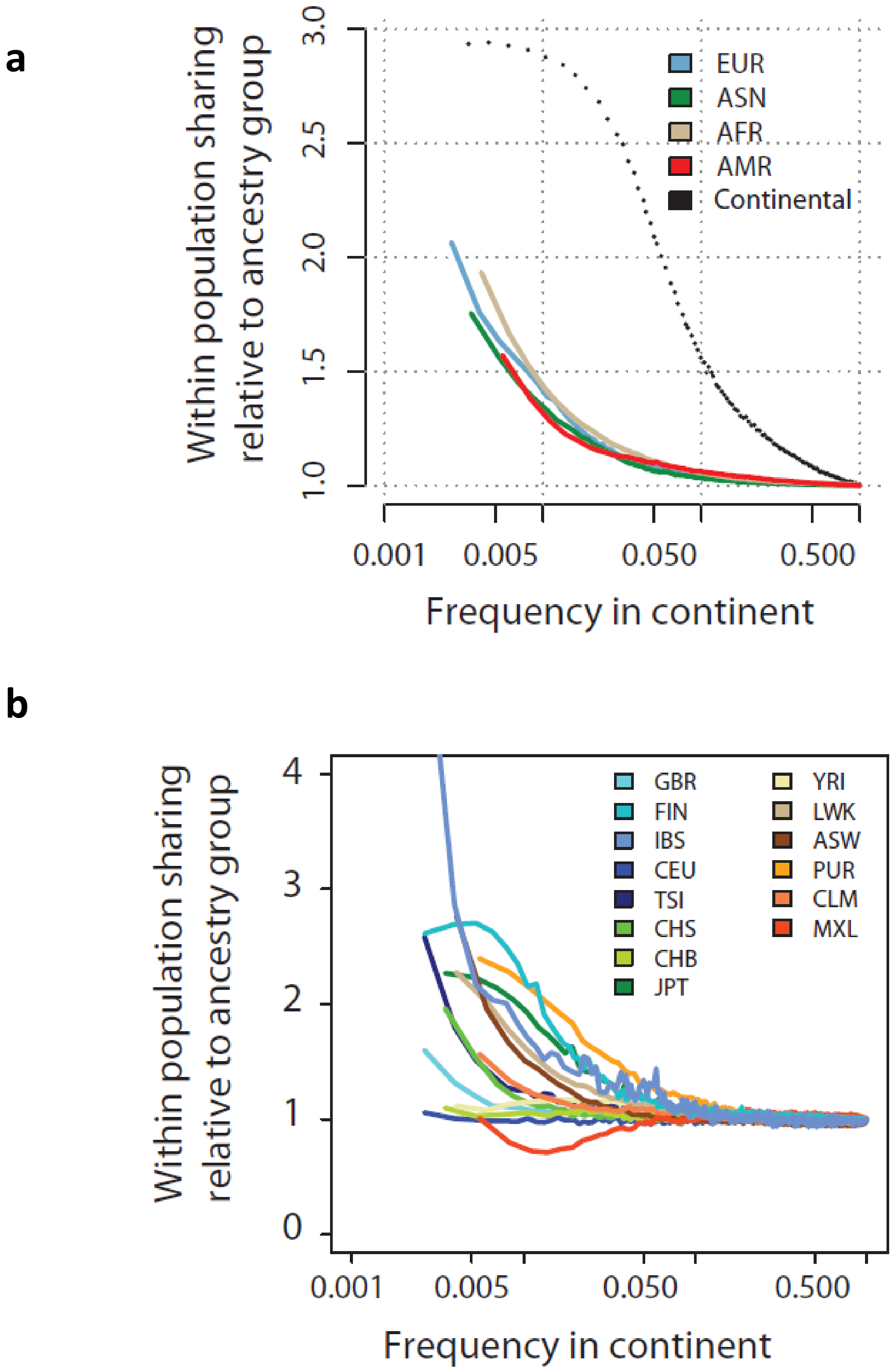


Figure S6. Geographical differentiation of rare variants.

a, Excess within-population (compared to wider ancestry-based grouping – see Figure 2a for a definition of which populations are in which group) allele sharing as a function of variant frequency within the group. Metric defined as the ratio of the probability of picking (without replacement) two chromosomes that share a variant within a population (weighted by the number of pairs within each population) compared to the same probability across the wider group (see Supplement). Dotted line indicates the excess within-continent (ancestry-group) sharing. **b**, As for part a, but the excess within-population sharing metric is calculated separately for each population within its ancestry-based group. The statistic for MXL samples drops below 1 for variants between 0.5% and 5%, indicating a relative dearth of variants in this allele frequency range (across the ancestry group) within the population.

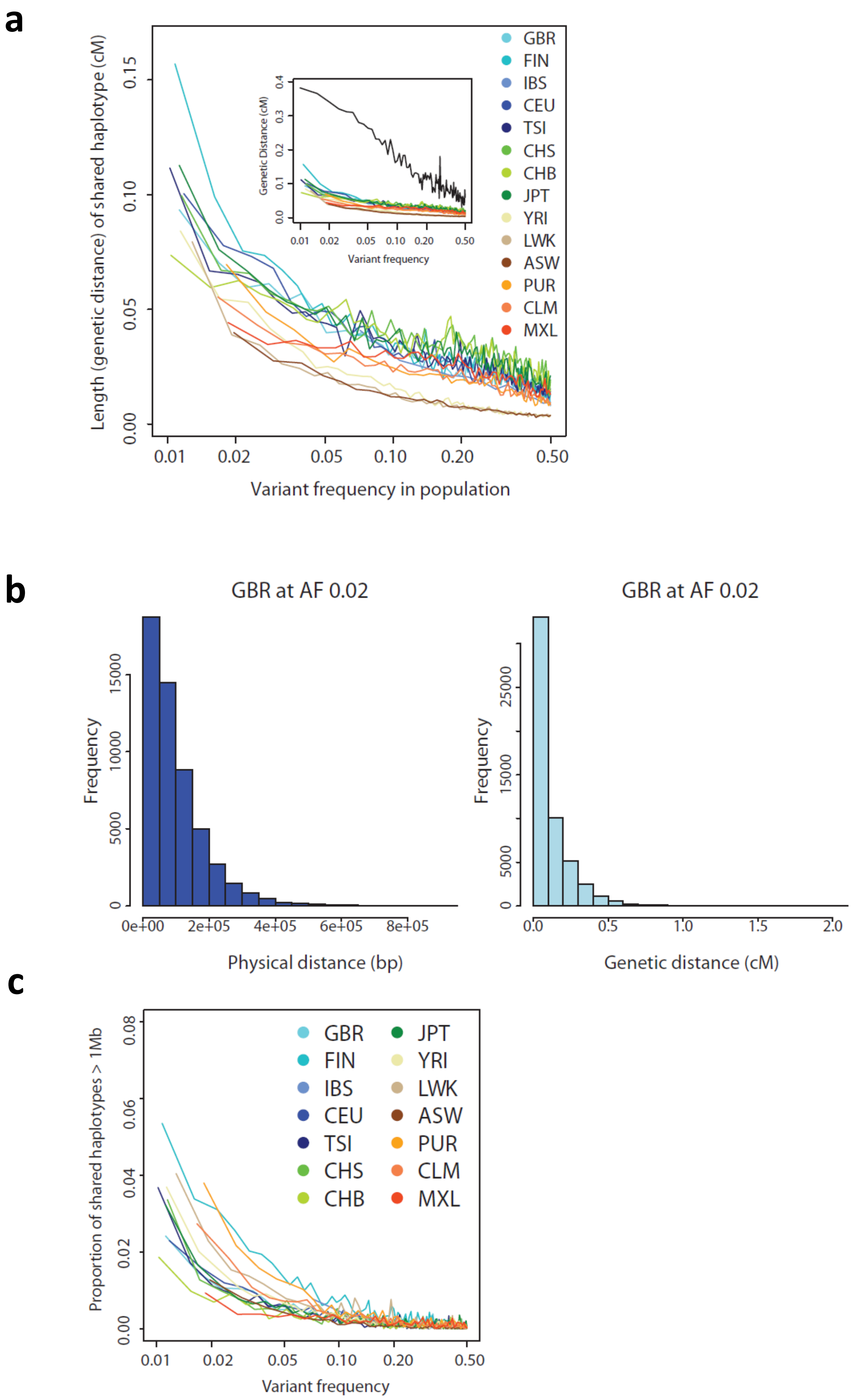


Figure S7. The length of shared haplotypes around variants of different frequencies.
a, Median genetic length of shared haplotype identity for pairs of chromosomes carrying variants of different frequency in each population (removing cryptically-related samples, singleton variants and allowing for up to two genotyping errors). The inset shows the expectation from a model of explosive recent population growth (Nelson et al. 2012) in which an effective population size of 10,000 has grown to 4 million in the last 10,000 years (assuming a generation time of 25 years). **b**, The distribution of physical (left) and genetic (right) shared haplotype lengths for variants of frequency 2% in the GBR population. **c**, The fraction of shared haplotypes that extend over 1Mb as a function of variant frequency in each population.

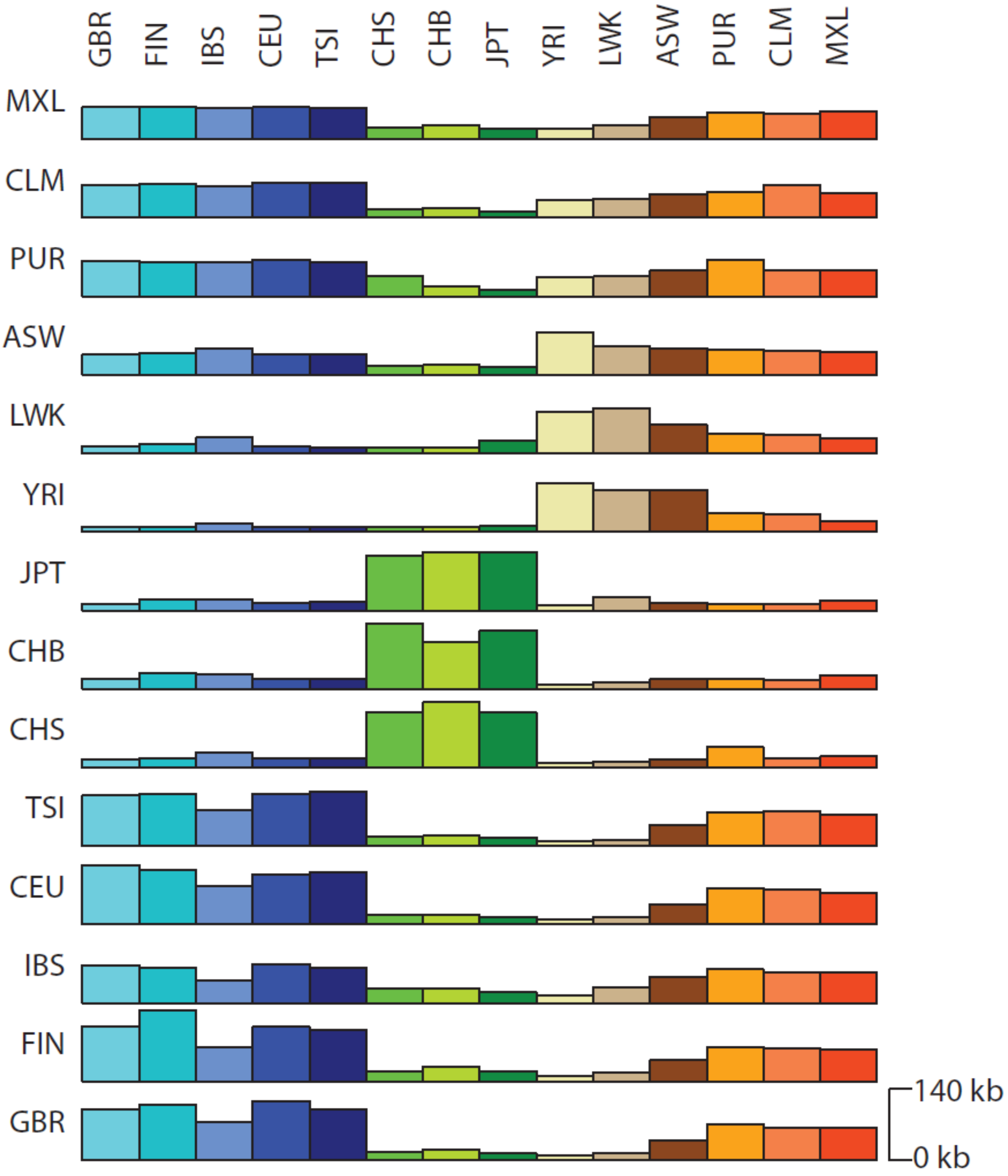
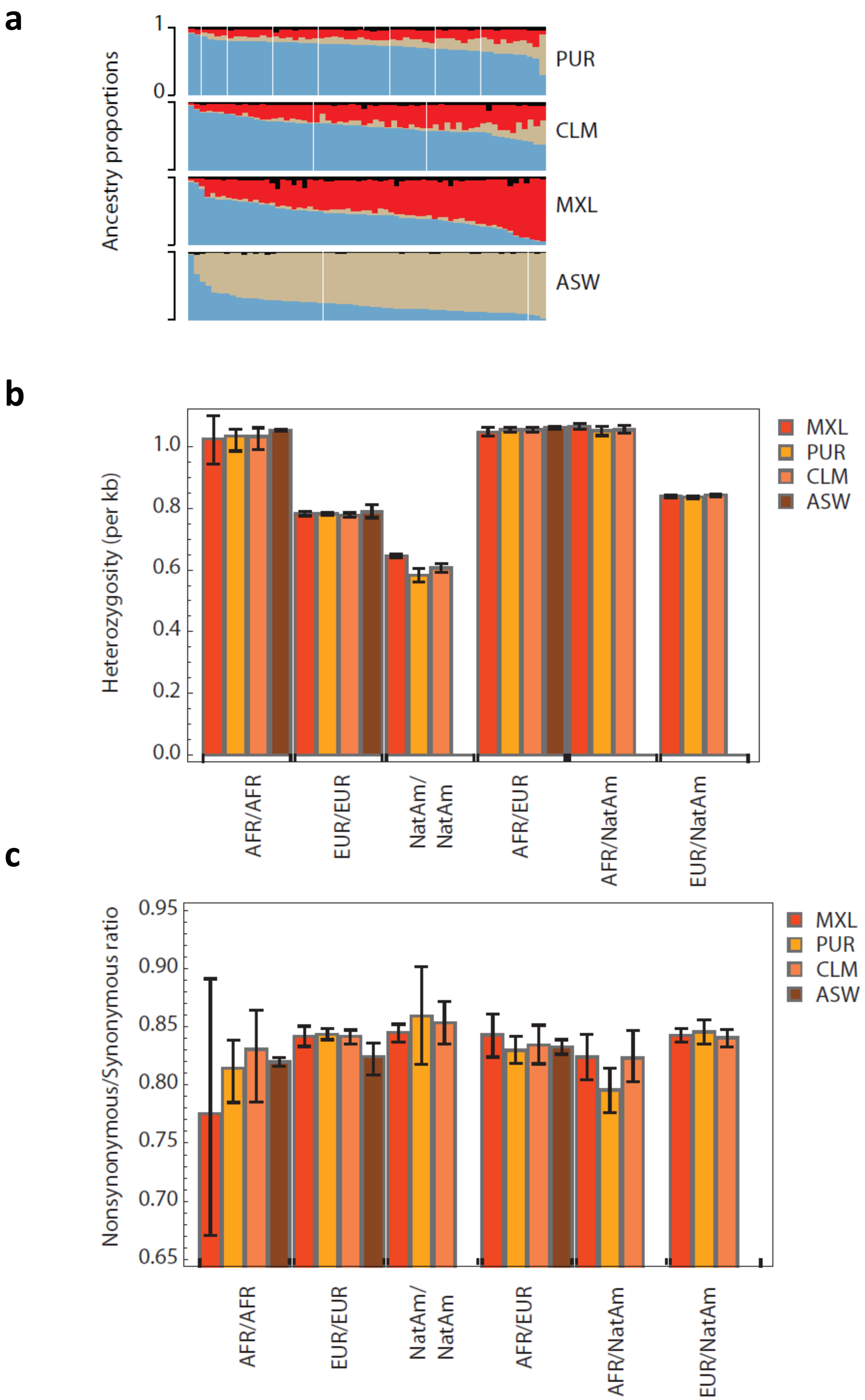


Figure S8. Shared haplotype length around f2 variants shared within and between populations. Summary of median physical length of haplotypes around variants present exactly twice across the sample, broken down by the population origin of the chromosomes sharing the variant. Bar heights are normalised to the maximum across the graph (within FIN; 140 kb).



Figures S9. Properties of genetic variation in regions of different inferred ancestry within the populations sampled from the Americas.

a, Estimated ancestry proportions for individuals in the ASW, MXL, CLM and PUR populations (blue: European, light-brown: African, red: Native American, black: unassigned regions). **b**, Average per base heterozygosity. **c**, Ratio of nonsynonymous to synonymous variants within the same regions. Error bars estimated from bootstrap re-sampling.

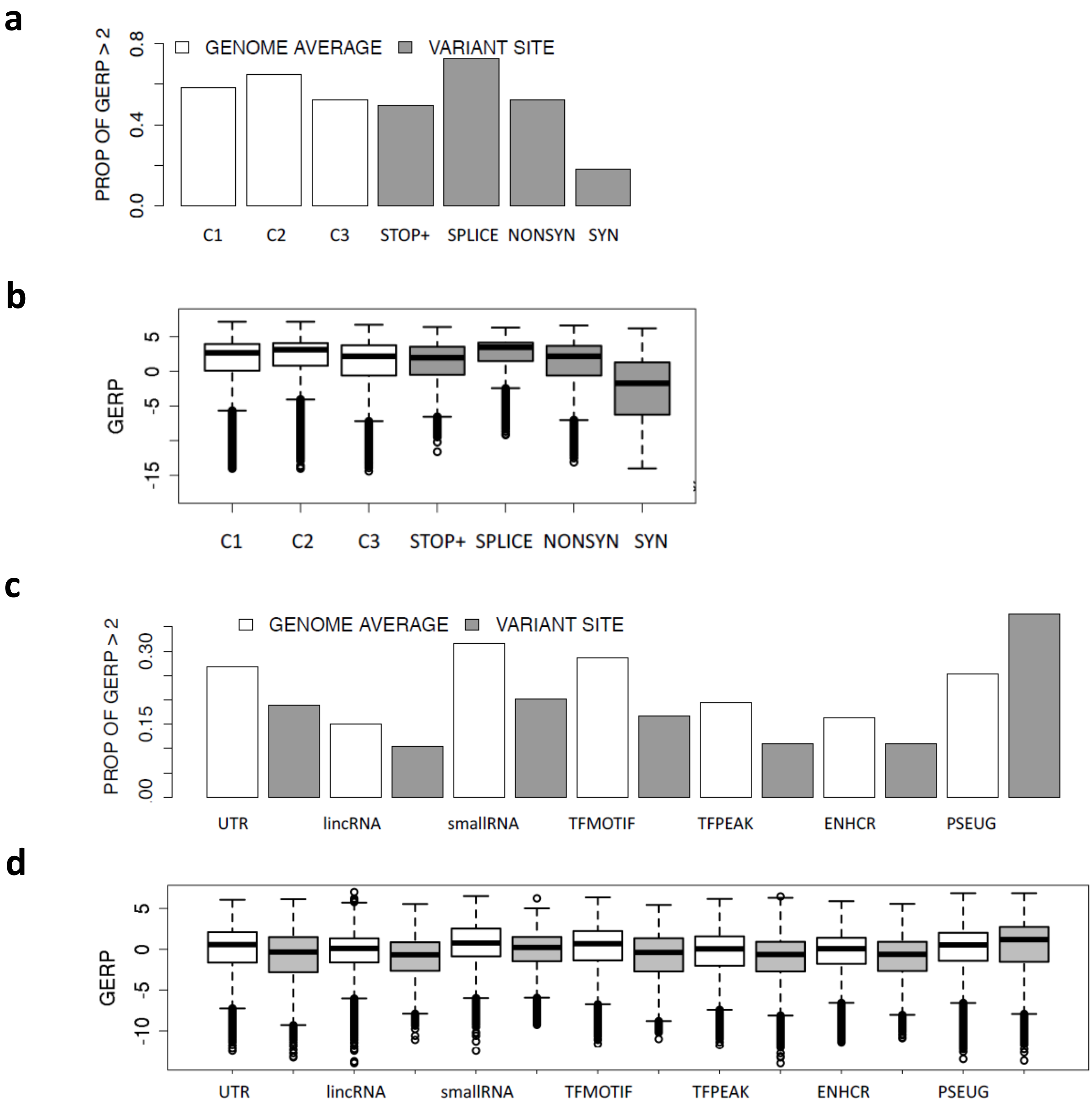


Figure S10. Conservation and variation by sequence annotation and variant type.
a, The fraction of sites at each codon position (C1-C3), and at sites showing different types of variant, where the evolutionary conservation (GERP) score is greater than 2. **b**, Boxplots showing the distribution of GERP scores within part A. Note that GERP scores are a function of the site, not the variant type. **c**, The fraction of sites within noncoding features of different types at all sites (white) and sites showing variation (grey). **d**,. Boxplots showing the distribution of scores for part c. Apart from pseudogenes, those sites at which variation is observed show a consistently lower level of conservation than the class as a whole.

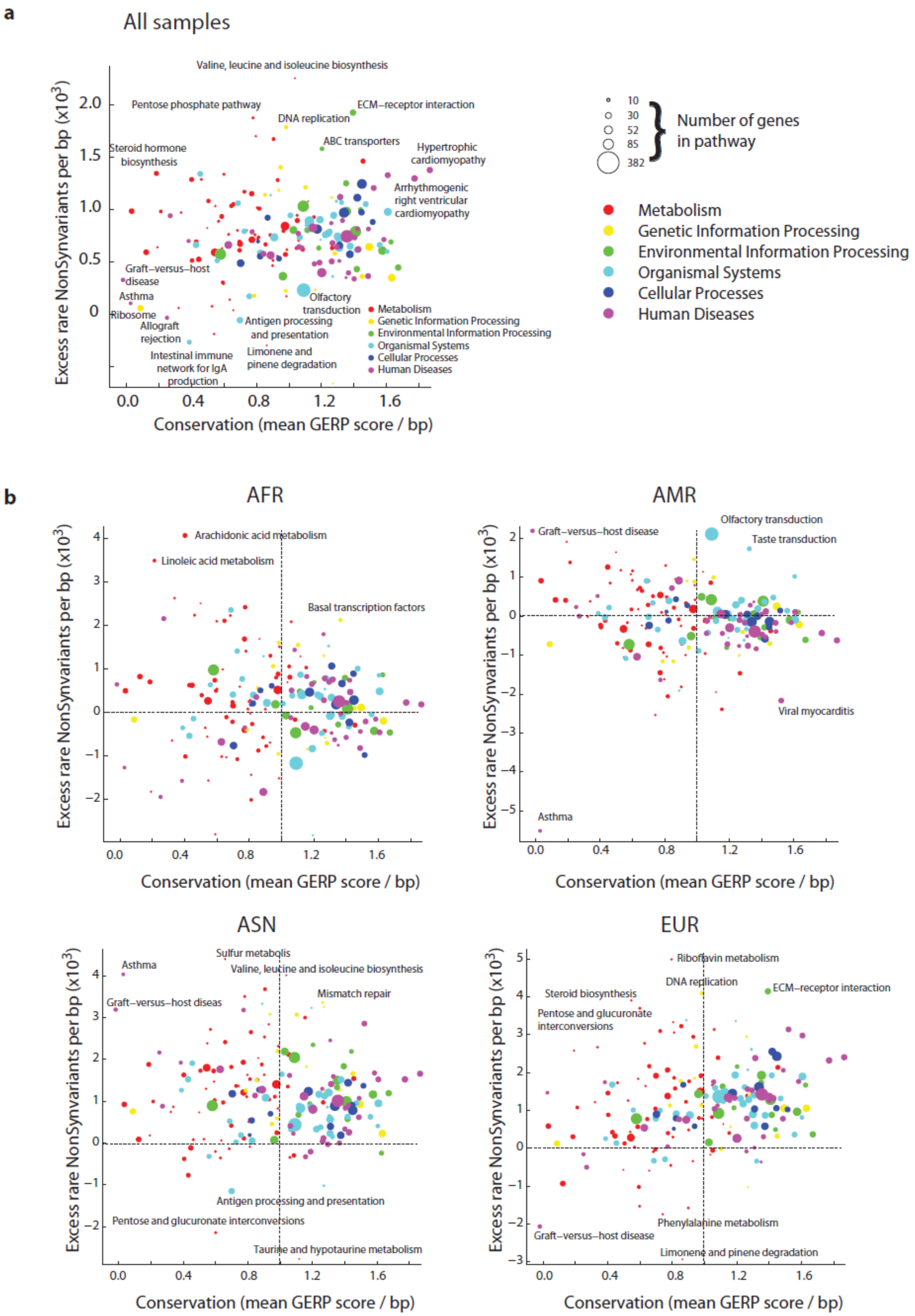


Figure S11. Excess rare nonsynonymous variants by KEGG pathway.

a, The relationship between evolutionary conservation and load of rare NonSyn variants in KEGG pathways estimated from , where *N* and *S* are the number of NonSyn and Syn variants across a pathway respectively and R and C represent rare (<0.5%) and common (≥0.5%) variants respectively. Negative values arise from a higher NonSyn:Syn ratio among common than rare variants. Dot area is proportional to the number of genes in the KEGG pathway. **b,** Excess rare nonsynonymous mutations in KEGG gene pathways for each ancestry-based group of populations (defined as in Figure 2A). Selected KEGG pathways are identified.

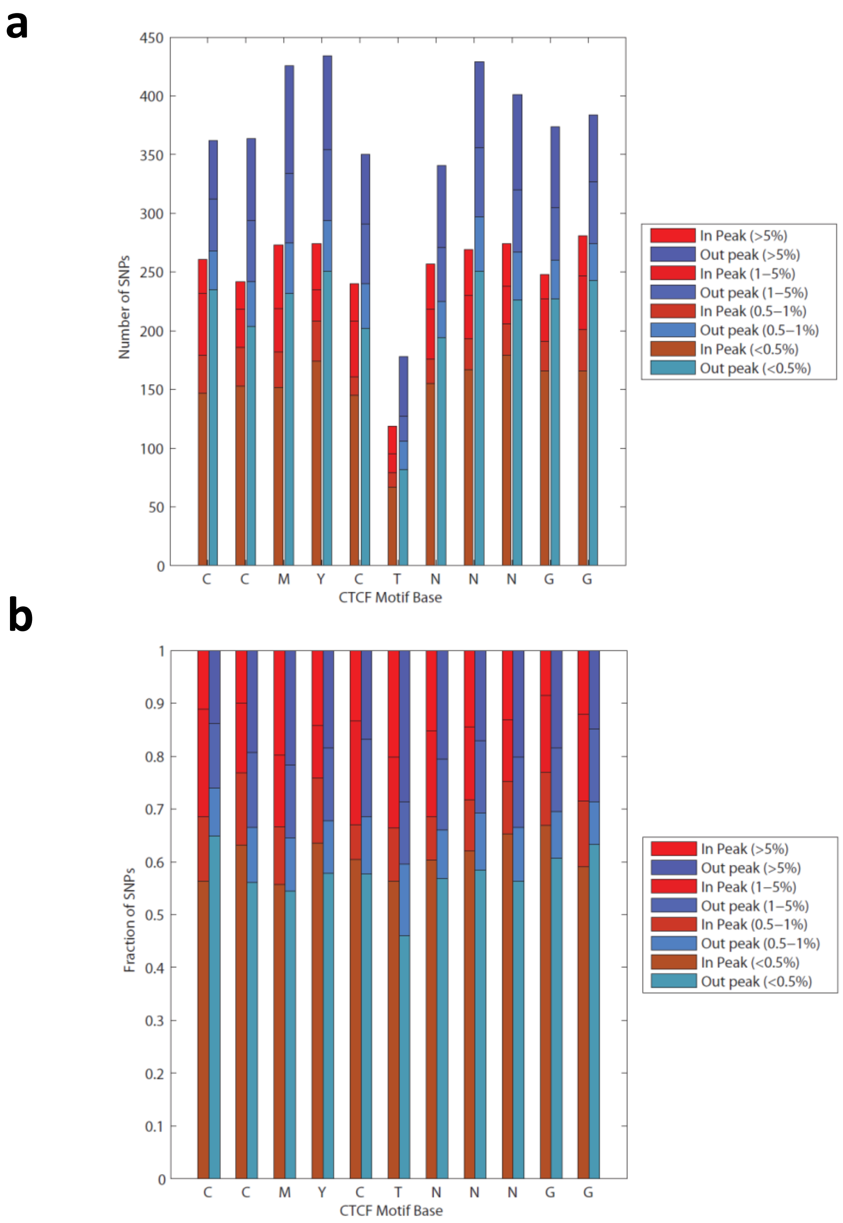


Figure S12. Analysis of SNP density and allele frequency around CTCF motifs.
SNP density and allele frequencies around the CTCF-binding motifs shown in Figure 4c. **a**, SNP density stratified by global allele frequency. **b** Fraction of SNPs in each frequency class.

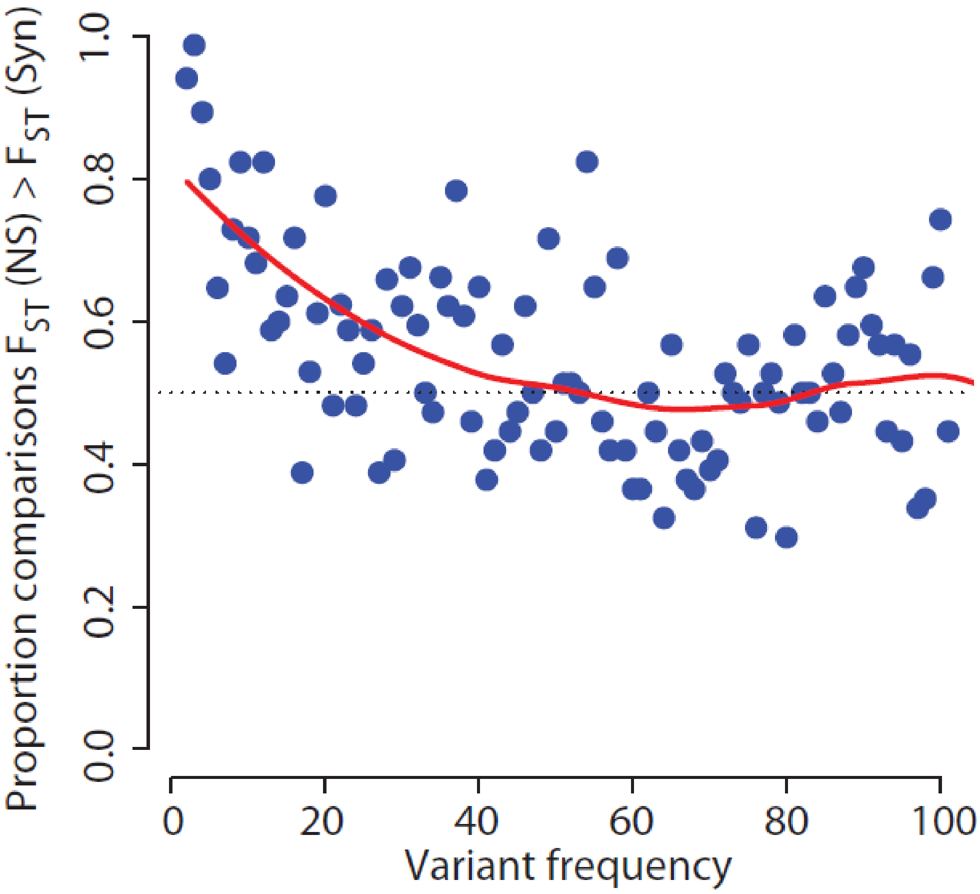


Figure S13. Difference between nonsynonymous and synonymous variants in population differentiation.

The fraction of pairwise population comparisons where nonsynonymous variants show greater differentiation, as measured by F_{ST}, than synonymous variants, at each allele frequency. The red line shows a smoothed estimate to highlight the trend.

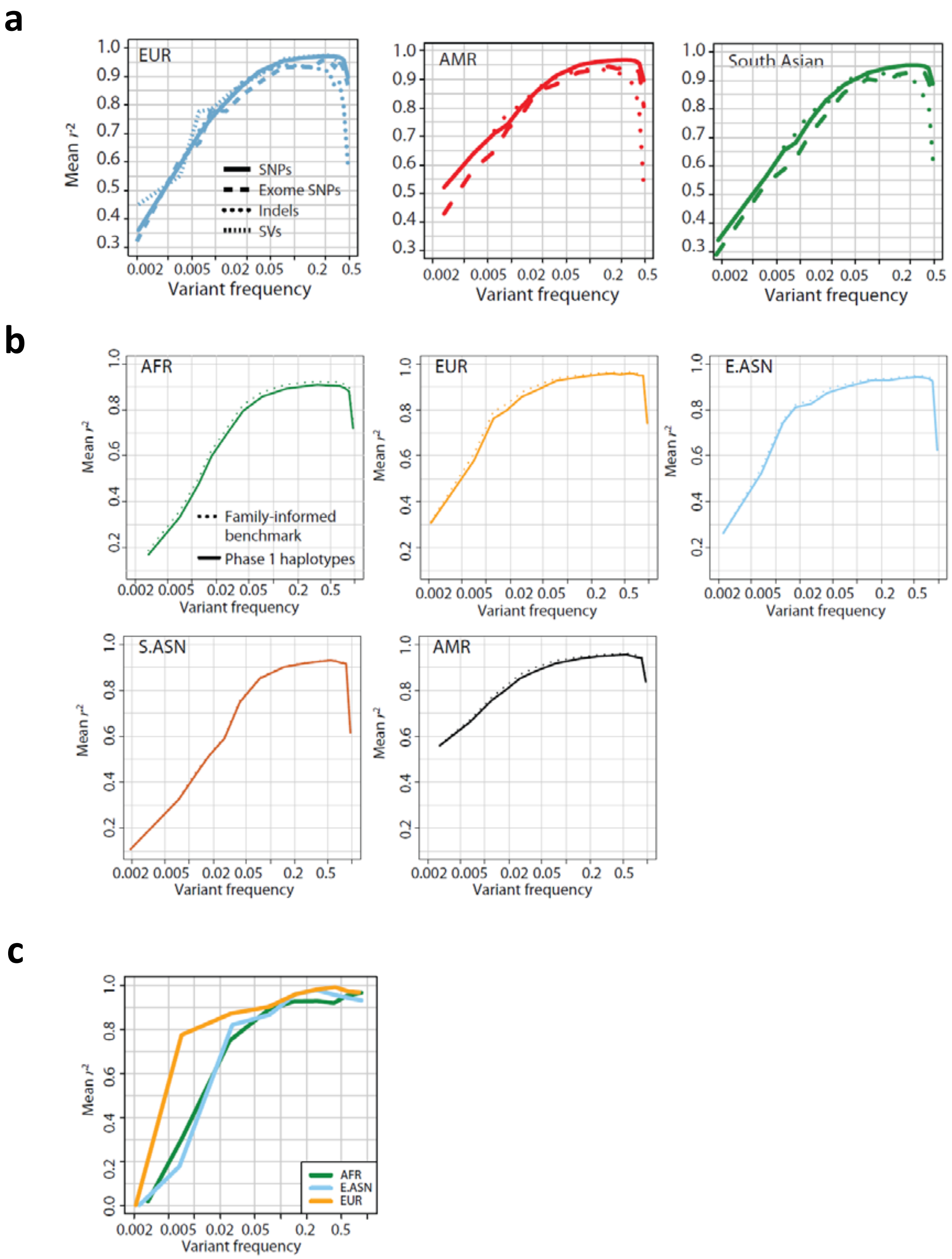


Figure S14 Accuracy of variant imputation.

a, Accuracy of imputation into 3 individuals from AMR (MXL), 4 individuals of European ancestry (3 CEU and 1 TSI) and 4 individuals of South Asian ancestry (Gujarati from Houston: GIH). Lines as for Fig. 5b. None of the imputed samples were sequenced in the current phase of the project.

b, Comparison of imputation of high quality SNP genotypes from a backbone of haplotypes estimated using family information (a mixture of duos and trios) to imputation from the Phase 1 release haplotypes, as a function of variant frequency. In each population group imputation from the Phase 1 haplotypes is only slightly worse than from the benchmark data, indicating that variant frequency and haplotype structure are the primary determinants of imputation performance. **c**, Accuracy of large deletion imputation for samples within Phase 1 arising from the fact that SV genotype likelihoods were only calculated for samples with Illumina sequencing data (see Supplement for details). Concordance measured against genotypes from Conrad et al. (2006).

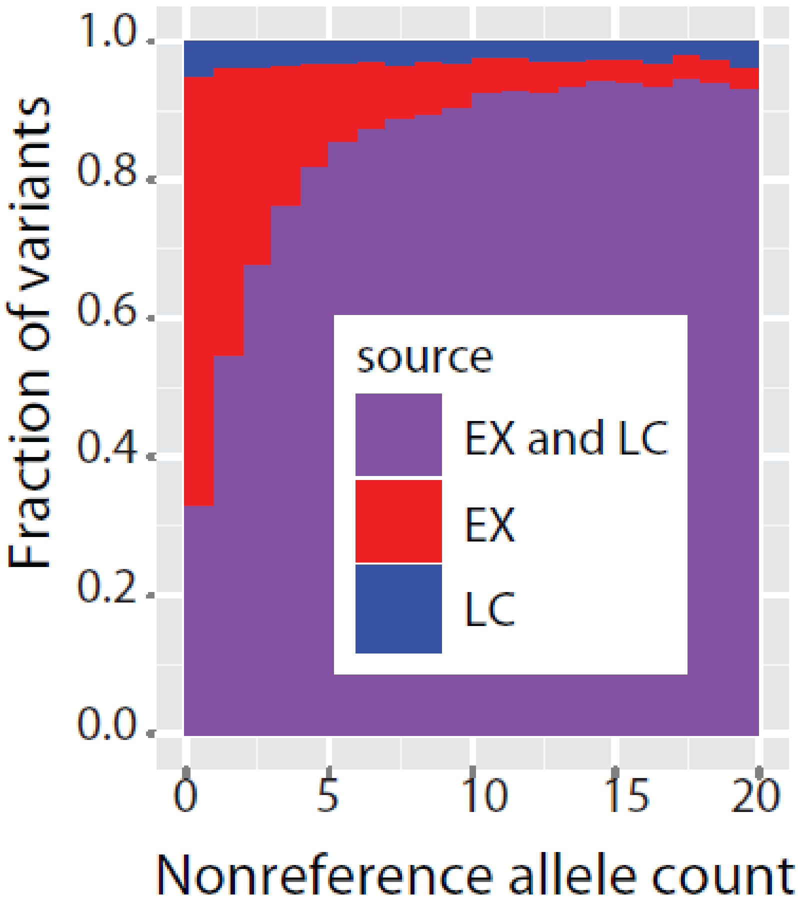


Figure S15. SNP discovery from low-coverage and targeted exome data.
Within the target exome consensus (24.3 Mb spanning 194,041 exons of 15,412 genes; see Supplementary Information), the plot shows the fraction of SNPs that were identified from both low-coverage and exome data (purple), exome-data only (red) and low-coverage data only (blue) as a function of the estimated variant count from the integrated haplotype release. About 60% of singleton variants were detected only from the exome data. At higher frequencies, about 10% of SNPs are discovered using only one approach, reflecting differences in the processing, analysis and filtering of variants from the different data sources. These will reflect a mixture of true and false positives from each approach. Details on the consensus target for the exome analysis can be found at:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/exome_pull_down/.