doi:10.1038/nature15393

Contents

1 Introduction

In this Supplementary Text we give further technical information regarding the 1000 Genomes Phase 2 and 3 data collection, processing, validation, and analysis. The aim is to record in more detail than is possible in the main text how the callsets were generated and analysed. The 1000 Genomes Phase 3 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. Corresponding authors for individual sections are underlined.

Extended Data Figure 1 shows a summary outline of the steps and datasets that went into creating the Phase 3 call set. Boxes indicate the relevant section of the supplement to refer for more details.

2 Materials

2.1 Description of Samples

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

2.1.1 Choice of populations included in the project

The population sampling plan for the project was based on the goal of finding 95% of common variants, looking at a broad set of continental backgrounds. As explained in the pilot paper[1], an efficient way to find rare variants is to sample a set of geographically related populations with about 1% F_{ST} differentiation among them. For variants that were rare in the original ancestral population, genetic drift in the set of related populations could raise or lower the frequencies of rare variants, generally different ones in different populations, so that the probability of finding each variant in at least one local population would be enhanced. As such, studying five related populations with 100 samples each should be more effective at finding rare variants than sampling one population with 500 samples[1].

The 1000 Genomes Project studied a total of 2,504 samples, about 500 samples from each of five continental ancestry groups, with generally five populations for each group. For the samples with ancestry from Europe, East Asia, and South Asia, populations across the geographic range had about 1% F_{ST}. The amount of divergence among African populations is much higher; it would require sampling many more populations to capture rare variation adequately. With a limited budget, the project decided to focus on populations related to the Yoruba, but not to attempt to be comprehensive within Africa. Other projects, such as the African Genome Variation Project[2] and the Human Health and Heredity in Africa (H3Africa) Project[3] are studying variation in additional African populations. For populations in the Americas, known admixture among groups with ancestry from Europe, Africa, and the Americas led the project to adopt a different strategy. Two populations had

primarily African and European ancestry (ASW from the U.S. and ACB from the Caribbean) and four populations (MXL, CLM, PUR, PEL) with a wide range of European, African, and indigenous American ancestry chosen to represent the wide variation in ancestry proportions observed in North, Central, and South America.

The project Steering Committee made the final decisions about which populations to include in the project, based on recommendations by the Samples and ELSI Group, according to the criteria listed in the Phase 1 paper[4] and outlined below. Members of the Samples and ELSI Group had expertise in population genetics, bioethics, and social science. One or more representatives from each sample collection team also participated, bringing specific expertise in the populations being studied.

The samples studied in Phase 1 were those initially available for sequencing and represented 14 populations. Data from the full set of samples, representing 26 populations, were used in this paper with the populations listed in Supplementary Information Table 1.

The project tried to include samples from mother-father-adult offspring trios wherever possible, but for some populations it was infeasible to obtain a sample from more than one person in a family. Information about the number of trio samples per population is available at `http://www.1000genomes.org/about#ProjectSamples`. The offspring in trios were not part of the 2,504 samples sequenced as part of the main dataset, but were often genotyped using a high-density microarray (Section 3.4), or sequenced using Complete Genomics (Section 3.5).

2.1.2 Other considerations guiding the selection of populations to be included

Several other considerations guided the recommendations of the Samples and ELSI Group about which population samples to include. The goal was to focus on relatively large, major populations rather than small, indigenous isolates. This is consistent with the project's goals being primarily medical rather than anthropological. It also reflects a recognition that members of small isolates may be more vulnerable to identification, invasion of privacy, stigmatization, or other harms.

The Samples and ELSI Group determined that, within these constraints, it was unnecessary for any particular population to be studied. The preference was to collect in the country of each population's ancestry, but this was not always feasible, so some populations sampled lived far from their ancestral locations. Most of the specific populations selected for inclusion were chosen because their samples had been collected with broad consent and had been studied extensively (e.g., the HapMap

samples) or because of well-established researcher-community connections in particular locales. Sampling in most cases was conducted in collaboration with a research centre or hospital where a relationship of trust had already been established and where other studies, including GWAS, were being done. This minimized the risks of misunderstanding of the research and increased the likelihood that the data will provide some scientific benefit for the studied populations, such as improving the interpretation of genetic studies in those populations.

The Samples and ELSI Group decided that it was appropriate to include in the project already existing samples collected for the HapMap without obtaining new informed consent. The consent form used for the HapMap contained broad language about placing data on millions of genetic variants in people on the internet. The Samples and ELSI group decided that this language was sufficient, even though it did not use the term "sequencing". The group was influenced in this determination by the fact that with the exception of the CEU, no traditional identifiers or medical information were collected with any of the HapMap samples (and for the CEU, all the phenotype information and links to the donors' identities are held in confidence by the local investigators, thus mitigating potential privacy concerns). Existing sets of samples were approved for use by the project in a couple of other cases. However, in each case, such approval was given only after a rigorous assessment of the adequacy of the original consent (using the criteria described below), and in some cases, it was required that the donors be re-consented specifically for this project. In these cases, it was also required that all traditional identifying and phenotype information be maintained in confidence.

2.1.3 Population labels

The criteria for assessing whether a prospective sample donor qualified as a "member" of particular population for purposes of the project varied from site to site to take into account local circumstances. In most cases, however, prospective donors were required to have at least 3 out of 4 grandparents who identified themselves as members of the group being sampled. Additional information about how population membership was assessed for each population can be found at the Coriell web site `https://catalog.coriell.org/1/NHGRI/Collections/1000-Genomes-Collections/1000-Genomes-Project`.

Another issue is the choice of the label adopted for each population (and its associated three-letter abbreviation). A number of scientific, cultural, and practical issues were taken into account in deciding on the label to assign to each population. It is important that investigators who use project data or samples in their studies use the

officially-sanctioned population labels consistently, both for scientific precision and to minimize the risks associated with over-generalization of research findings. The samples studied for the project are not meant to be perfectly representative of all people in the specific location where the sampling occurred, all people in the general geographic region, or all people with ancestry from that region.

Depending on the scientific question being asked, it may be appropriate to pool data from a particular sample set with data from samples collected from other ancestrally related groups. For example, if the groups all have African ancestry, the designation "African ancestry" to describe the combined analysis panel is recommended. Guidelines for how to label the populations included in this project, and for the rationale behind the approach to labelling, can be found at `https://catalog.coriell.org/1/NHGRI/About/Guidelines-for-Referring-to-Populations`.

2.1.4 Informed consent criteria

The Samples and ELSI Group developed a set of criteria for use by project investigators proposing to collect new samples for the project or to obtain new consent from participants in previous genetic studies; these criteria were also used to assess the adequacy of existing consent forms. Consent forms were required to state explicitly that:

- Extensive individual data from the study of the samples (but no individual identifiers or medical information) would be made publicly available in scientific databases on the Internet;
- Samples and data would be labeled by population and comparisons among individuals and populations would be made;
- Individual samples could not be withdrawn, and once data from the study of samples had been put in the database, the data could not be withdrawn;
- Individuals or communities would not have an opportunity to pre-approve future uses of the samples, although all proposed uses would be assessed by the staff of the Coriell Institute, where the samples are stored, to ensure that the samples will be used only in ways that are consistent with the terms of the informed consent;
- Cell lines would be made, making it possible to generate a potentially unlimited amount of DNA that may last indefinitely;

- Samples and data would be made available to many researchers around the world;
- Samples and data would be used not only for the 1000 Genomes Project, but also for many other future projects (including gene expression studies, studies of population history and relatedness, etc.);
- Samples and data would be used by academic, commercial, and government entities, and if such uses resulted in the development of commercially valuable products, participants would not share in the proceeds;
- No individual results from the study of the samples would be returned, although general information about the project and about interesting new findings emerging from genetic research is provided periodically to the researchers who collected the samples, and who are encouraged to share them with community members.

The Samples and ELSI Group developed a consent form template that incorporates the above criteria, which is available at `http://www.1000genomes.org/sites/1000genomes.org/files/docs/Informed%20Consent%20Form%20Template.pdf`. The template was intended as a starting point for investigators but was subject to modification, depending on local requirements and cultural norms, as long as the core requirements were kept.

In addition to incorporating these conditions, all consent forms had to comply with the laws and regulations in the country where the samples were to be collected. In some cases, it was also necessary to secure licenses from the appropriate regulatory authorities to have samples transported outside the country.

2.1.5 Review and approval of sampling plans

The Samples and ELSI Group required the PI for each sample collection team to submit a written sampling plan for review, discussion, and approval. These plans provided members of the Samples and ELSI Group with basic information about the protocols to be used for sample collection at each site so that the group could ensure that uniform standards were followed across all sampling sites (taking into account local requirements and cultural norms), could identify any major issues likely to arise, and could provide guidance about possible ways to address these issues.

Investigators were asked to address these issues in their sampling plans:

- The social and demographic characteristics of the population and the communities where samples were to be collected (e.g., the size of the population and the communities), the socio-economic and educational status of members of the communities, and any unique features of the population that might be relevant to the project (e.g., minority status, history of discrimination, previous participation in biomedical research).
- The process to be used to obtain informed consent (e.g., how researchers would explain what the project was about, how they would ensure that people did not feel pressured to participate, plans on translating consent materials into the local language, and plans on reading the consent form orally to people who could not read).
- The processes they would use to identify and address any concerns that might be relevant to the broader population or communities (e.g., consulting community leaders, holding public meetings, or using other means to seek group input or responses to group concerns).
- Any particular concerns they anticipated might arise in the community, and how they proposed to address them. These included concerns about:
  - The physical process of drawing blood.
  - The cultural or symbolic meaning of drawing blood.
  - Genetic research in general (e.g., concerns about making cell lines or about the possibility of human "cloning").
  - Privacy and the potential for individuals to be identified.
  - Use of the samples or data for commercial purposes.
  - Having the samples sent to a repository in the U.S.
  - Having the data in a database maintained by a U.S. government health research agency (NIH).
  - The possibility that the research results could be used to stigmatize or discriminate against individuals, the communities, or the population.
  - The possibility that participating in the research could have a negative impact on the family (where trios were to be collected).
- What the researcher considered to be the best way to inform the donor community about the progress of the project and about how their samples would be used in the future.

When concerns of the type mentioned in the written sampling plans came up in the communities (which happened infrequently), the Samples and ELSI Group discussed the issues and suggested ways to resolve them.

2.1.6 Sample distribution and follow-up with participating communities

All blood samples collected for the project were sent to the National Human Genome Research Institute (NHGRI) Repository for Human Genetic Research at the non-profit Coriell Institute for Medical Research in Camden, New Jersey, which transformed them into lymphoblastoid cell lines. The Coriell Institute provides each research group that collected samples with a free set of DNA or cell lines from those samples. Researchers in resource-limited countries are charged a substantially reduced rate for sample DNA.

The Coriell Institute publishes its inventory of DNA and cell lines through an online catalogue (`https://catalog.coriell.org/1/NHGRI`). As a privacy safeguard, more samples were collected from each population than were actually studied for the project; it is thus unknown whether the sample from a particular person ended up being used for the project. All researchers who order the samples must submit a written Statement of Research Intent that describes the nature of the research they plan to conduct, so that the Coriell Institute can determine whether the proposed research is consistent with the terms of the informed consent signed by the sample donors. The Coriell Institute provides regular reports to the principal investigators of the sample collection groups that describe how the samples they collected, and data from those samples, are being used. The PIs are encouraged to share these reports with the participating communities.

The Coriell Institute regularly solicits feedback from members of the participating communities regarding any issues or concerns they may have about the project or about how their stored samples are being used. So far, no concerns have been raised.

2.2 Lymphoblastoid cell line establishment and quality assurance

Authors: Neda Gharani, Lorraine H. Toji and Norman P. Gerry

Lymphoblastoid cell cultures were established at the Coriell Cell Repositories from fresh bloods after separating the mononuclear cells (PBMCs) on a Ficoll gradient and incubating with Epstein Barr virus and phytohemaglutinin in RPMI 1650 with

15% v/v fetal bovine serum[5].

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[6]. 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[7] 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 Promega PowerPlex® 18D System to resolve ambiguities) and an amelogenin gender assay; these data are also used to confirm family relationships of trios.

To ensure that no cryptic (unexpected) first and second degree relationships were present among members of a given population, an additional quality assurance step was added for samples collected as part of the final population collection phase (BEB, ESN, GWD, ITU,MSL, PJL and STU). This step involved Affymetrix 6.0 SNP Array genotyping of DNA samples from viable cell lines that had passed all initial sterility and identify QCs. Identity by state (IBS) data analysis was carried out using the Partek Genomic Suite, which includes an IBS analysis module[8]. Cryptic relatedness in all the other populations was identified by post hoc IBS analysis using Illumina Omni2.5 or HapMap SNP data (analysis carried out by Dr. James Nemesh at the Broad Institute). Replacement samples were subsequently identified by Coriell and added to the project.

For the ESN population, frozen PBMCs isolated from fresh blood at the University of Ibadan, Nigeria prior to shipping, were submitted to Coriell for cell line transformation. From three other 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.

Details of the available DNA and cell lines can be found on the 1000 Genomes website (`http://www.1000genomes.org/cell-lines-and-dna-coriell`).

3 Data generation and processing

3.1 Reuse of data from the Pilot and Phase 1 data sets

Authors: Laura Clarke, Xiangqun Zheng-Bradley

For the final stage of the project, the consortium decided to use only Illumina platform sequence and data from reads which were 70 bp or longer. This means that not all data produced for the Pilot[1] and Phase 1[4] were used in the final Phase 3 dataset.

As the majority of the Pilot sequencing was carried out before read lengths reached 70 bp, very little of the Pilot data could be reused; only 99/12185 runs from 17/742 different samples were retained. A larger quantity of the Phase 1 data could be reused, although sequence data generated using the SOLiD and LS454 platforms or with read lengths shorter than 70 bp was removed. For the low coverage sequencing 3721/13774 runs were retained from 798/1396 samples. For the exome sequencing, 9045/9702 runs were retained from 816/1128 samples. The larger number of retained exome runs reflects the fact that the majority of Illumina platform exome sequencing in the first phase was carried out using $\geq$70 bp read lengths. Due to these changes, there are 59/1092 samples from Phase 1 that are not found in Phase 3, due to either insufficient retained sequence data, or the sample not having both exome and low coverage data.

3.2 Low-coverage whole genome and whole exome sequencing

3.2.1 Broad Institute

Sequencing: Low-coverage and exome
Authors: Namrata Gupta, Stacey Gabriel and The Broad Institute's Genomics Platform

Starting with 250 ng or less of input DNA, samples are quantified using a PicoGreen assay and diluted to a working stock volume and concentration, then libraries are constructed and sequenced on Illumina HiSeq 2000 or HiSeq 2500 with the use of 76 bp or 101 bp paired-end reads. Output from Illumina software is processed by the Picard data-processing pipeline to yield BAM files containing well-calibrated, aligned reads. All process steps are performed using automated liquid handling instruments, and all sample information tracking is performed by automated LIMS messaging.

Library Construction

Libraries are constructed using the protocol described in Fisher *et al.*[9] with several modifications: first, initial genomic DNA input into shearing has been reduced from 3 μg to 100 ng in 50 μL of solution. Second, for adapter ligation, Illumina paired end adapters have been replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter. These index sequences enable pooling of libraries prior to sequencing. Third, custom sample preparation kits from Kapa Biosciences were used for all enzymatic steps of the library construction process.

In-solution hybrid selection (for whole exome libraries)

In-solution hybrid selection was performed as described by Fisher *et al.*[9].

Size selection (for whole genome shotgun libraries)

Size selection was performed using Sage's Pippin Prep, with a target insert size of either 340 bp or 370 bp $\pm$ 10%. Multiple gel cuts were taken for libraries that required high (>8×) sequencing coverage.

Preparation of libraries for cluster amplification and sequencing (whole exome libraries)

Following in-solution hybrid selection, libraries are quantified using PicoGreen. Based on PicoGreen quantification, libraries are normalised to equal concentration and pooled by equal volume. Library pools are then quantified using a Sybr Green-based qPCR assay, with PCR primers complementary to the ends of the adapters (kit purchased from Kapa Biosciences). After qPCR quantification, library pools are normalised to 2 nM, denatured using 0.2 N NaOH, and diluted to 20 pM, the working concentration for downstream cluster amplification and sequencing. After running an aliquot of the denatured library pool on an 8-cycle Illumina MiSeq run (only 8 bp indices are sequenced) to verify pool evenness, pools are spread across the number of flowcell lanes needed to meet target coverage for all samples within the pool. Uneven pools are re-pooled before cluster amplification sequencing.

Preparation of libraries for cluster amplification and sequencing (whole genome libraries)

Following size selection, libraries are quantified using a Sybr Green-based qPCR

assay, with PCR primers complementary to the ends of the adapters (kit purchased from Kapa Biosciences). After qPCR quantification, libraries are normalised to 2 nM, denatured using 0.2 N NaOH, and diluted to 20 pM, the working concentration for downstream cluster amplification and sequencing.

Cluster amplification and sequencing
Cluster amplification and sequencing of denatured templates are performed according to the manufacturer's protocol (Illumina) using v3 cluster amplification kits, v3 flowcells, v3 Sequencing-by-Synthesis kits, Multiplexing Sequencing Primer kits, and the latest version of Illumina's RTA software. Whole exome runs are 76 bp paired-end on either HiSeq 2000 or HiSeq 2500, and whole genome runs are 101 bp paired-end on HiSeq 2000.

3.2.2 Baylor College of Medicine – Human Genome Sequencing Center

Sequencing: Low-coverage and exome
Authors: Michelle Bellair, Huyen Dinh, Harsha Doddapaneni, Viktoriya Korchina, Christie Kovar, Donna Muzny

Library Construction
DNA samples were constructed into Illumina paired-end libraries according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications as described in the BCM-HGSC protocol[10]. Libraries were prepared using Beckman robotic workstations (Biomek NXp and FXp models). Briefly, 1 μg of genomic DNA was sheared into fragments of approximately 300-400 base pairs with the Covaris E210 system, followed by end-repair, A-tailing, and ligation of the Illumina multiplexing PE adaptors. Pre-capture ligation-mediated PCR (LM-PCR) was performed for 7 cycles of amplification using the Phusion PCR Supermix HiFi (2X) (NEB, Cat. No. M0531L). Purification was performed with 1.8X Agencourt AMPure XP beads (Beckman, Cat. No. A63882) after enzymatic reactions, and following the final purification, quantification and size distribution of the pre-capture LM-PCR product was determined using the LabChip GX electrophoresis system (PerkinElmer).

Capture Enrichment
Six uniquely barcoded pre-capture libraries were pooled together in equimolar amounts (totalling 2 μg per pool) and then hybridised in solution to the HGSC VCRome 2.1 design1 (42Mb, NimbleGen) according to the manufacturer's protocol NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2) with minor revisions. Human COT1 DNA and full-length blocking oligonucleotides were added into the hybridisation to block repetitive genomic sequences and the adaptor sequences.

Post-capture LM-PCR amplification was performed using the Phusion PCR Supermix HiFi (2X) with 12 cycles of amplification. After the final AMPure XP bead purification, quantity and size of the capture library was analysed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The efficiency of the capture was evaluated by performing a qPCR-based quality check on the four standard NimbleGen internal controls. Successful enrichment of the capture libraries was estimated to range from a 6 to 9 of ΔCt value over the non-enriched samples.

Sequencing
Library templates were prepared for sequencing using Illumina's cBot cluster generation system with TruSeq PE Cluster Generation Kits (Cat. No. PE-401-3001, PE-402-4001) according to the manufacturer's protocol. Briefly, these libraries were denatured with sodium hydroxide and diluted to 6-9 pM in hybridisation buffer in order to achieve a load density of ~800K clusters/mm2. Each library pool was loaded in a single lane of a HiSeq flow cell, and each lane was spiked with 2% phiX control library for run quality control. The sample libraries then underwent bridge amplification to form clonal clusters, followed by hybridisation with the sequencing primer. Sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 and 2500 platforms. Using the TruSeq SBS Kits (Cat. No. FC-401-3001, FC-402-4001), sequencing-by-synthesis reactions were extended for 101 cycles from each end, with an additional 7 cycles for the index read. Sequencing runs generated approximately 300-460 million successful reads on each lane of a flow cell, yielding ~6 Gb per sample. With these sequencing yields, samples achieved an average of 89% of the targeted exome bases covered to a depth of 20× or greater.

Whole genome sequencing
Libraries prepared for whole genome sequencing followed the library construction protocol described above, with the following modifications: DNA was sheared into fragments approximately 500-700 base pairs, and 0.8X AMPure XP beads were used for purification of the fragmented DNA. Sequencing also generally followed the methods above, with the following exceptions: three uniquely barcoded libraries were pooled, and then the library pools were loaded in two lanes of a HiSeq flow cell to achieve the desired coverage. With this approach, samples achieved an average of ~12.2 Gb per sample (~8.2× coverage).

3.2.3 BGI

Sequencing: Low-coverage and exome
Authors: Xiaosen Guo, Tianming Lan, Bo Wang, Xuedi Ma, Jun Wang

For whole-genome sequencing, we constructed DNA libraries according to Illumina

recommended protocols. a) We randomly fragmented 3-5 μg genomic DNA to less than 800 bp by Covaris E210/LE220/S2. b) We performed the end repair by trimming the 5' overhangs and filling the 3' overhangs by use of the T4 DNA polymerase, Klenow fragment, T4 PNK and dNTPs. c) In order to add adaptors at both ends of DNA fragments, we processed the blunt DNA fragments to ligate an "A" at 3' ends. d) Ligated adaptors to both ends of DNA fragments by using T4 DNA Ligase. f) We then performed an agarose gel electrophoresis with a concentration of 2% to separate DNA products, and DNA fragments with a length between 450 and 550 bp were recycled and purified according to the user guide of Qiagen Gel Extraction Kit. g) Performed a PCR enrichment to ensure that we have enough DNA products to be successfully sequenced, and the primer we used in this step were 1.1 and 2.1 (Illumina). Finally, these DNA fragments were subjected to the Illumina Hiseq 2000 platform for pair-end sequencing. We used the Illumina Base Calling programs (BclConverter-1.9.0 or OLB-1.9.4) to convert the image files to sequence files with the data format of FASTQ. The read length was 90 bp.

For exome capture sequencing of Phase 2 and 3, we used the NimbleGen capture array from SeqCap EZ Human Exome Library kit of Roche NimbleGen to enrich exonic DNAs of each library. We first randomly fragmented the genomic DNA to 200-250 bp by Covaris E210/LE220/S2, and the adaptors were ligated to both ends of each fragment by using DNA ligase. After several cycles of ligation-mediated PCR reaction, the DNA fragments with the sequencing adapters were then hybridised to the NimbleGen capture array following the protocol. Then we eluted the enriched DNA fragments from the array and washed out DNA fragments which were not hybridised to the array. The enriched DNA was amplified by PCR. And then these enriched DNA fragments were subjected to the Illumina HiSeq 2000 platform. Each captured library was sequenced independently and 90 bp reads were finally generated by converting the raw image files through the same Illumina Base Calling software with default settings.

3.2.4 Max Planck Institute for Molecular Genetics

Sequencing: Low-coverage
Authors: Tatiana A. Borodina, Ralf Herwig, Hans Lehrach, Ralf Sudbrak, Bernd Timmermann, Vyacheslav S. Amstislavskiy, Matthias Lienhard, Marcus W. Albrecht, Marc Sultan, Marie-Laure Yaspo

Genomic DNA sequencing of samples for Phase 3 of the 1000 Genomes project was fulfilled between October 2012 and March 2013 on HiSeq2000 (Illumina).

Genomic DNA was obtained from the Coriell Institute for Medical Research. Li-

braries were prepared starting from 1 μg of genomic DNA. DNA was fragmented by ultrasound (Covaris S220) to obtain fragments in the 300-400 bp range. Fragmented DNA was further processed into paired-end (PE) libraries using TruSeq DNA Sample Prep v2 Low Throughput Protocol. 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 10 nM library stocks before loading them onto flowcell to reach optimal cluster density) to make the process more reproducible and predictable. Sequencing was performed on Illumina HiSeq2000 platform using paired-end 2×101 bp sequencing mode allowing identification of 101 nucleotides from each side of the genomic DNA insert of library molecules. All libraries were barcoded which allowed more than one barcoded sample per sequencing lane. The barcoding indexes were determined using additional 6 bp sequencing read according to Illumina's instructions.

Raw data was pipelined according to corresponding manufacturer's instructions. Illumina's bcl2fastq software v1.8.2 was used for base calling and demultiplexing. For preliminary analysis, resulting sequencing reads were aligned to the human genome (hg18, NCBI build 36.1).

3.2.5 Washington University

Sequencing: Low-coverage and exome
Authors: Richard Wilson, Elaine Mardis, Li Ding, Lucinda Fulton, Bob Fulton, Dave Larson, Laura Courtney

All 1000G Phase 3 samples were received as DNA samples from the Coriell Cell Repository. For each sample selected for whole genome sequencing, we prepared an indexed whole genome library. These indexed libraries were pooled together. Sequence generation used an Illumina HiSeq 2000 instrument with 100 bp paired end runs and all samples were sequenced to 4× coverage. For each sample selected for whole exome sequencing, we prepared an indexed library. These libraries were pooled and enriched using Nimblegen SeqCap EZ Human Exome Library v.3.0 following the manufacturer's instructions. Sequence generation used an Illumina HiSeq 2000 instrument with 100 bp paired end runs. Each sample was sequenced to a depth of >20× coverage over >80% of the targeted region.

3.2.6 Wellcome Trust Sanger Institute

Sequencing: Low-coverage
Author: Anja Kolb-Kokocinski

Whole Genome Sequencing: All of the genomic DNA was obtained from Coriell Institute for Medical Research. Samples have a basic quality check performed to assess suitability for submission to sequencing or genotyping. The volume is checked using a BioMicroLab automated volume check system. Pico Green Assessment of concentration is performed using both Beckman FX liquid handling platforms and Molecular Devices plate readers. Invitrogen E-Gels are run to check sample integrity; the loading of these gels is automated using Beckman FX/NX liquid handling platforms. A standard Sequenom assay containing 26 autosomal and 4 gender markers is performed to produce a fingerprint of the samples, which is used to confirm identity post-sequencing or genotyping. The gender markers also allow for sample swaps and plate orientation issues to be identified prior to downstream analysis.

Whole Genome Sequencing: For library generation genomic DNA (approximately 1 μg) was fragmented to an average size of 500 bp and subjected to DNA library creation using established Illumina paired-end protocols. Adapter-ligated libraries were amplified and indexed via PCR. A portion of each library was used to create an equimolar pool comprising 8 indexed libraries. Libraries were subjected to 100 base paired-end sequencing (HiSeq 2000; Illumina) following manufacturer's instructions.

3.2.7 Illumina

Sequencing: Low-coverage
Authors: David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury

GenomicDNA was acquired from the Coriell Institute for Medical Research. Paired end Illumina sequencing libraries were prepared using the TruSeq PCRFree method[11] starting with 500 ng of material resulting in libraries with average insert size of 382 bp. Libraries were denatured using NaOH (0.1 N) and diluted in cold (4 °C) hybridisation buffer prior to seeding clusters on the surface of the flow cell. Cluster amplification, linearisation, blocking and hybridisation to the Read1 sequencing primer were carried out on a cBOT. Following the first sequencing read, flow cells were held in situ and clusters were prepared for Read2 sequencing. Paired end sequence reads of 101 bases were generated using the HiSeq2000, as described in the

Illumina Instrument guide[12]. A total of 90% of PF reads had a raw read accuracy of ≥Q30.

3.3 High-coverage whole genome PCR-free sequencing

Author: Namrata Gupta

For comparison to the low-coverage sequencing in the larger sample, we selected a number of individuals for high-coverage sequencing using a PCR-free protocol. High-coverage sequencing was performed for a CEU trio, a YRI trio, and one individual from each of the remaining 24 populations. Specifically, the following individuals were selected: HG01879 - ACB, NA19625 - ASW, HG03006 - BEB, HG00759 - CDX, NA12892 - CEU, NA12891 - CEU, NA12878 - CEU, NA18525 - CHB, HG00419 - CHS, HG01112 - CLM, HG02922 - ESN, HG00268 - FIN, HG00096 - GBR, NA20845 - GIH, HG02568 - GWD, HG01500 - IBS, HG03742 - ITU, NA18939 - JPT, HG01595 - KHV, NA19017 - LWK, HG03052 - MSL, NA19648 - MXL, HG01565 - PEL, HG01583 - PJL, HG01051 - PUR, HG03642 - STU, NA20502 - TSI, NA19238 - YRI, NA19239 - YRI, and NA19240 - YRI.

Illumina PCR-free fragment shotgun libraries were prepared using the 'with-bead pond library' construction protocol described by Fisher *et al.*[9] with the following modifications. 500 ng of genomic DNA, in a volume of 50 μl, was sheared to a size of ∼400 bp using a Covaris E210 instrument (Covaris) using Illumina's TruSeq PCR-free protocol shearing parameters (Illumina, Part # 15036187 A): Duty cycle = 10%, intensity = 5, cycles per burst = 200, time = 45 seconds. Fragmented DNA was then cleaned up with 0.6× Agencourt AmPure XP SPRI beads and eluted in 40 μl Tris-HCl pH8.0, following manufacturer's recommendations (Beckman Coulter). DNA fragments were then further clean up with 3.0x Agencourt AmPure XP SPRI beads, following manufacturer's recommendations (Beckman Coulter), but DNA was not eluted from the SPRI beads. Then using the KAPA Library Preparation Kit reagents (KAPA Biosystems, Catalog # KK8241) DNA fragments bound to the SPRI beads were subjected to end repair, A-base tailing and adapter ligation using Illumina's 'PCR-free' TruSeq adapter (Illumina, Catalog FC-121-3001) following manufacturer's recommendations (KAPA Biosystems). A second 0.7x SPRI reaction was performed after adapter ligation in lieu of a size selection step to tighten up size distribution and eliminate any excess adapters. No library PCR amplification enrichment was performed. Sequence ready Illumina PCR-free library was then eluted off the SPRI beads in 25 μl of Tris-HCl pH8.0 following manufacturer's recommendations (Beckman Coulter). Libraries were quantified with quantitative PCR using KAPA Library Quant kit (KAPA Biosystems, Catalog # KK4824) and

an Agilent Bioanalyzer High Sensitivity Chip (Agilent Technologies) following the manufacturer's recommendations.

Libraries were sequenced with 250 base paired-end reads on an Illumina HiSeq2500 instrument in Rapid Run Mode, with the following modifications. Reagents from two 200 cycle TruSeq Rapid SBS Kit v1 (Illumina, catalogue # FC-402-4001) were combined and run using a 500 cycle run. To enable a 500 cycle run the `<SBSMAXCycleRR>` value in the HiSeqControlSoftware.Options.cfg file was changed to 500 cycles i.e. `<SBSMAXCycleRR>500</SBSMaxCycleRR>`. According to Illumina it is also possible to define the number of cycles in the HiSeq Control Software under the Run Configuration tab, however entering non-supported read length will result in a warning message. Currently Illumina does not support read lengths greater than 2×150 bases on the HiSeq2500 with the v1 chemistry, however they plan to do so with the next release.

3.4 High-density microarray genotype data

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip`

3.4.1 OMNI, Broad Institute

Authors: George Grant, Wendy Brodeur

Illumina Infinium Genotyping Protocols: The Illumina Infinium pipeline for whole genome genotyping uses two Biomek F/X, three Tecan Freedom Evos, and two Tecan Genesis Workstation 150s to automate all liquid handling steps throughout the sample and chip preparation process. The entire process of reagent addition and sample manipulation is completed without any manual steps to alleviate or minimise any batch effects or human error. Samples begin the process in 96 well plates at a concentration of 20 ng/μl after being qualified for the particular chip type during the pre-genotyping QC steps, which will be described in the sample handling section below. First, samples undergo a whole genome amplification step; four micro-litres (80 ng) is transferred to a plate so that they can be denatured and neutralised. Amplification occurs overnight in an oven. After amplification, the samples are enzymatically fragmented using end-point fragmentation. The next two steps, precipitation and resuspension, are to clean up the DNA before hybridisation onto the chips. The precipitation step simultaneously allows for an in-process QC step. The precipitation occurs with a reagent that adheres to the DNA and has a

blue colour so that technicians can visually affirm that there is adequate amplified DNA to proceed through the rest of the steps.

The passing fragmented, resuspended DNA samples are then dispensed onto the appropriate BeadChips and placed in the hybridisation oven to incubate overnight for 22 hours. After hybridisation, the chips are washed to remove unhybridised and non-specifically hybridised DNA sample from the BeadChips. Labeled nucleotides are added to extend the primers hybridised to the DNA by one base and the primers are immediately stained and the Beadchips are coated for protection before scanning. The BeadChips are scanned with one of the four Illumina iScan Readers with dual autoloaders, which use a laser to excite the fluorophore of the single-base extension product on the beads. The scanner records high-resolution images of the light emitted from the fluorophores. All plates and chips are barcoded and tracked with an internally derived laboratory information management system (LIMS); this allows confidence that the correct sample was applied to the appropriate chip.

Illumina Infinium Quality Control Procedures: We have established a number of valuable platform-specific QC checks. After every chip is detected in production genotyping, and the genotypes transferred to the database, a series of chip-level QC checks are performed to determine if the chip data is acceptable to be passed on to analysis:

Fingerprint concordance: Depending on when the project was complete, we either ran samples with our historic Sequenom panel, or the newer Fluidigm Fingerprinting panel (both described below). Samples may or may not have both assay calls.

Sequenom: All samples intended for Illumina production genotyping are fingerprinted using our Sequenom MassArray platform prior to being run on the Illumina whole genome arrays. Samples are genotyped on Sequenom using a specially designed fingerprint panel that includes 24 SNPs that are also on the Illumina array and one gender assay. Prior to plating samples for Illumina, the fingerprint genders are compared to the genders reported to the Broad by the collaborators. The results of the gender comparison are reported to the collaborator. This gender comparison can be an opportunity to detect labelling or arraying errors via gender discordance, and also provides some indication of DNA suitability for Infinium. Concerning samples can be eliminated at this point in the process.

Fluidigm: The Fluidigm fingerprint panel includes 29 SNPs that overlap with the Affy 6.0 array and have multiple proxy SNPs each, 66 SNPs that overlap with Illumina's 1m and 2.5m arrays and have multiple proxy SNPs each, 32 SNPs in transcribed regions of housekeeping genes that are expressed in most cell types and 1 gender determining SNP. The results of the gender comparison are reported to the

collaborator. This gender comparison can be an opportunity to detect labelling or arraying errors via gender discordance, and also provides some indication of DNA suitability for Infinium. Concerning samples can be eliminated at this point in the process.

After Illumina Infinium genotyping is completed, a genotype concordance check is performed to see if the genotypes for the Fingerprint Panel SNPs on the array match the fingerprint. If the sample is not >90% concordant with the fingerprint, it is not passed. This test permits robust determination that the Illumina genotypes are derived from the correct sample and are not due to contamination with another sample or other PCR products. This analysis can identify potential sample swaps. Any potential sample mix-ups undergo further investigation and are reported to the collaborator. If a sample swap is confirmed, that is the Illumina genotype matches the fingerprint of a different sample, then an reassignment of the sample genotype is possible. Comparison of the fingerprint genotype with the Illumina genotype for a given sample allows us to ensure that the correct sample and genotype are associated with one another.

QC Data cleaning: Each Illumina whole genome genotyping project undergoes a rigorous, manual data review process that ultimately leads to the re-calling of each SNP genotype using a project-specific custom cluster file. Once 75% of the project is completed in the lab we begin the QC data cleaning process using Illumina's BeadStudio software. Initially, genotypes are called using Illumina's Autocall cluster file before our trained technical staff QCs the data manually.

For autosomal chromosomes, our team reviews poor performing SNPs and "zeros" out SNPs if any of the following occur:

- GenTrain score of less than 0.6 (the GenTrain score is a QC metric that Illumina uses to measure how well a SNP clusters at a given position).
- Cluster Separation score of less than 0.4.
- Number of replicate errors greater than 2.
- Number of parent-parent-child errors greater than 1.

The sex chromosomes undergo an independent manual review process. We manually review X, Y, XY, XX and mitochondrial SNPs. SNPs are "zero'd" out if any of the following occur:

- When anomalies are observed such as undefined clusters, aberrant intensity normalisation, or abnormal cluster positions.

- Adjustment of genotype cluster centres with simultaneous re-calling of the genotype as needed.

After poor performing SNPs are removed, the project-specific cluster file is generated and the genotypes are 're-called' using this custom cluster file.

Sample call rate: For Illumina production genotyping, a sample must have a call rate greater than or equal to 98% after being called using the chip specific/project specific cluster file to be considered passing. Samples with a call rate <98% are considered to have failed first-pass genotyping. Failed samples are reported to the collaborator, who is given the option to re-run those samples that have enough DNA available. Samples that fail due to a processing error will be re-run at no cost to the collaborator. Any samples that fail due to unknown causes can be re-run at the collaborator's expense.

HapMap concordance: All plates bound for Illumina Infinium whole genome genotyping include a HapMap process control sample, which is one of the samples genotyped during International Haplotype Map (HapMap) project. The Genomics Platform has standard HapMap trios from different populations that are available to plate with the experimental samples. Each of the three trio samples are rotated through the plates such that the trio ends up being repeated every three plates. The well position of the HapMap control is also rotated through the plates in a defined pattern, which can serve as a secondary plate identifier beyond the plate barcode. Rotation of the control sample avoids any location or inadvertent bias that could arise if the same process control were plated in the same location all of the time. The HapMap control on each plate serves as a control for all processing steps. We not only check the call rate of the HapMap controls but we compare the concordance of the called genotypes against the gold standard HapMap reference genotypes. If a plate is successful we expect that the HapMap controls should always have a call rate of greater than 98% and have a high HapMap concordance. Should either of these not be the case we investigate whether or not a processing error could have occurred while running that plate.

3.4.2 OMNI, Wellcome Trust Sanger Institute

Author: Anja Kolb-Kokocinski

Genotyping was performed using the Illumina Infinium BeadChip Omni2.5-8 following the manufacturers Infinium LCG Automated protocol.

Samples supplied to the institute are tested for quality and then quantified to

50 ng/μl by the onsite sample management team prior to submission to the Illumina Genotypying pipeline. Where possible, samples are submitted in plates of 96 or multiples of 8 to reduce array loss/costs. Before processing begins, manifests for submitted samples are uploaded to Illumina LIMS where each sample plate is assigned an identification batch so that it can be tracked throughout the whole process that follows.

Pre-Amplification: The Pre-Amplification process uses two Tecan Freedom Evos to process sample plates side by side allowing for six whole plates to be processed daily. The process is fully automated with the exception of a manual agitation/centrifugation step midway through and also at the end of the process. Four microlitres (200 ng) of sample is required (Illumina guidelines) for the Pre-Amplification reaction.

Post-Amplification: Over three days, Post-Amplification (Fragmentation, Precipitation, Resupension, Hybrisation to beadchip and xStaining) processes are completed as per Illumina protocol using four Tecan Freedom Evos. Following the staining process, beadchips are coated for protection and dried completely under vacuum before scanning commences on five Illumina iScans, four of which are paired with two Illumina Autloader 2.Xs.

Initial QC: Prior to downstream analysis, all samples undergo an initial QC to establish how successful the assay has performed. Sample call-rates below 92.5% are initially flagged before loading all samples into Illumina's GenomeStudio software. Using Illumina's QC dashboard, intensity graphs identify sample performance by measuring dependant and non-dependant controls that are manufactured onto each beadchip during production.

3.4.3 Affymetrix, Coriell

Authors: Neda Gharani, Norman P. Gerry, Lorraine H. Toji

Affymetrix genotyping followed the standard Affymetrix SNP 6.0 Protocol[13]. Briefly, 250 ng aliquots of genomic DNA were digested with either Nsp1 or Sty1. A universal adaptor oligonucleotide was then ligated to the digested DNAs. The ligated DNAs were diluted with water and three 10 μL aliquots from each well of the Sty 1 plate and four 10 μL aliquots from each well of the Nsp 1 plate were transferred to fresh 96-well plates. PCR master mix was added to each well and the reactions cycled as follows: 94 °C for 3 min; 30 cycles of 94 °C for 30 s, 60 °C for 45 s, 68 °C for 15 s; 68 °C for 7 min; 4 °C hold. Following PCR, the 7 reactions for each sample were combined and purified using Agencourt AMPure beads. The UV absorbance of the

purified PCR products was measured to insure a yield $\geq$4 μg/μL. 45 μL ($\geq$180 μg) of each PCR product was fragmented with DNAse 1 so the largest fragments were <185 bp. The fragmented PCR products were then end-labeled with a biotinylated nucleotide using terminal deoxynucleotidyl transferase.

For hybridisation, the end-labeled PCR products were combined with hybridisation cocktail, denatured at 95 °C for 10 min and incubated at 49 °C. 200 μL of each mixture was loaded on a GeneChip and hybridised overnight at 50 °C and 60 rpm. Following 16-18 hrs of hybridisation, the chips were washed and stained using the GenomeWideSNP6_450 fluidics protocol with the appropriate buffers and stains. Following washing and staining, the GeneChips were scanned on a GeneChip Scanner 3000. Genotype calls were generated from the scans of the arrays using Affymetrix Genotyping Console software which employs the birdseed2 genotyping algorithm.

3.5 Complete Genomics

3.5.1 Sample selection

Author: Lisa Brooks

Complete Genomics supplied deep sequencing data for 427 samples (see table below). Several considerations went into the choice of these samples. The project wanted sequence data for two trios from DNA for both blood and cell lines, to compare the data from the two types of DNA sources. The CEU and the YRI trios did not have both types of DNA available, but a trio from each from the PUR and KHV sample sets did. The Structural Variation Group chose 11 individual LWK samples for deep sequencing as a validation set for finding structural variants; this population was chosen because the samples were available at the time, some with DNA from blood (Buffy coats) and some with DNA from LCLs. A set of about 30 trios (with available samples, and with some differences from trio relatedness discovered after the samples were chosen) was sequenced from one population of each of the five continental groups (CEU, YRI, CHS, PEL, PJL), to provide good data for validation of the project results. 287 samples overlapped with samples in the primary low-coverage and exome data, while 284 overlapped with the samples in the final release.

Population	Blood/LCL	Samples	Trios	Duos	Unrelated	Male Probands	Female Probands	Overlap with Phase 3*
CEU	LCL	96	32			15	17	63
CHS	LCL	93	31			19	12	62
KHV*	LCL	3	1				1	3 (2)
KHV*	Blood	3	1				1	3 (2)
LWK	LCL	6			6			6
LWK	Blood (Buffy)	5			5			4
PEL	Blood	94	30	2		12	18	62
PJL	Blood	47	15	1		9	6	32
PUR*	LCL	3	1				1	3 (2)
PUR*	Blood	3	1				1	3 (2)
YRI	LCL	21	7			5	2	15 (14)
YRI	Blood (Buffy)	59	12	9	5	9	3	37
	Total	433	131	12	16	69	62	293 (288)
	Unique	427	129	12	16	69	60	287 (284)

3.5.2 CG data submission and processing pipelines

Authors: Christopher O'Sullivan, Chunlin Xiao, Bob Sanders, Shane Trask, Michael Kimelman, Eugene Yaschenko, Stephen Sherry

Complete Genomics delivered a data package distinct from those delivered to 1000 Genomes consortium for other NGS technologies. CG data packages are large compared to other NGS technologies, averaging 272 Gb per sample. The large size and unique structure of this data necessitated custom submission and processing pipelines for archiving raw data and delivering content to end users. The following information derived from the Complete Genomics data has been made available.

1. Alignment of evidence intervals to the reference in BAM format. Files are named "EvidenceOnly".
2. Alignment of evidence intervals to the reference with supporting reads in BAM format. Files are named "EvidenceSupporting".
3. Integrated VCF file containing sample genotypes for SNPs, indels, structural variants, and mobile insertion elements.
4. The Complete Genomics set of production reports as tarball. According to the Complete Genomics FAQ, `http://www.completegenomics.com/FAQs/Data-Results/`

Detailed documentation of the pipeline is available here:

`http://ftp.ncbi.nlm.nih.gov/sra/doc/SRA_CG_pipeline.pptx`
`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/cgi_variant_calls/NCBI_CG_data_flow_20150218.pdf`

SRA toolkit software:

`https://github.com/ncbi/sratoolkit/tree/master/tools/`

NCBI scripts used:

`http://ftp.ncbi.nlm.nih.gov/sra/utilities/`

Pointers to the CG BAM files on our FTP site can be found in the Complete Genomics indices directory

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/complete_genomics_indices`

Average coverage of the CG sequencing was ~55×. Basic alignment statistics for all runs are available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/cgi_variant_calls/NCBI_CG_summary_stats_20150219.txt`

3.5.3 Merged CG variant calls

Author: Goo Jun

Merged VCF files were created for the 427 samples based on the genotype files generated by the Complete Genomics pipeline. These merged files are available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/cgi_variant_calls/`

One set of files represent the short variants (SNPs, short INDELs and short substitutions). Low quality genotype calls were removed by cgatools varfilter, and the multi-sample VCF was generated by cgatools mkvcf. More information about cgatools is available at `http://cgatools.sourceforge.net`.

The other set of files represents the structural variations (large deletions) generated by multi-sample calling. Candidate intervals are collected from the Complete Genomics pipeline's single-sample CNV and SV junction intervals, and genotype calls were made by clustering sequencing depth across the 427 samples.

Filtered calls for evaluation:

The above files were further filtered for use in various evaluations. These filtered files are available here:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/cgi_variant_calls/filtered_calls/

The SNP and indel calls were subjected to this filtering.

1. All variants are normalized with left-aligned parsimonious form using 'vt normalize' tool available at http://genome.sph.umich.edu/wiki/vt
2. Variants with 10% or more missing genotypes were removed.
3. Variants with genotype-based inbreeding coefficient less than -0.1 were removed.

The SV files were filtered using this following process:

1. identification and merging of candidate intervals
2. multi-sample clustering and genotyping.

Candidate intervals were identified from junction events detected during read mapping and from sample-by-sample depth-based ploidy calls on every 2,000 base pair (bp) interval. Candidate intervals from the 1000G Phase 3 SV from low-coverage genomes were also included.

These candidates were then merged based on their reciprocal overlap. We next fitted Gaussian mixture models to sequencing depth profiles for each sample (after GC-correction) and evaluated clustering metrics to filter out ambiguous intervals.

Additional filtering was applied by removing all variants with Mendelian inconsistencies and HWE p-value less than 10^{-5}. SVs with >0.8 overlap with other SV with identical genotypes are also removed. The deletions callset was additionally filtered by Bayes Factor filtering.

3.6 Alignment

3.6.1 Decoy reference

Author: Heng Li

False read mapping is a source of false positive raw SNP calls[14]. These can often be removed by good filtering methods, but for Phase 3 we have tried to remove a common cause of these false mappings by including 'decoy' sequence into the reference genome used for read mapping. This decoy includes "novel human sequences"

that can capture reads from genuine human sequence that are not represented in the primary GRCh37 reference genome assembly. These "novel" sequences are frequently homologous to the chromosomal sequences in GRCh37. When absent, reads sequenced from the novel sequences will be mapped to the chromosomal sequences instead, which leads to spurious heterozygous calls showing unusual statistics such as high variant density, presence of three or more haplotypes at a locus, elevated read depth and violation of Hardy-Weinberg equilibrium.

Details of the construction of this decoy sequence are given in:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/README_human_reference_20110707`

The integrated reference sequence file (hs37d5.fa) used for Phase 3 mapping includes the GRCh37 primary assembly (chromosomal plus unlocalised and unplaced contigs), the rCRS mitochondrial sequence (AC:NC_012920), Human herpesvirus 4 type 1 (AC:NC_007605) and the concatenated decoy sequences described above. It is available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz`

As with the reference used in Phase 1, the pseudo-autosomal regions (PAR) on chromosome Y have been masked out by 'N's, so that the equivalent PAR on chromosome X may be treated as diploid for male samples. The coordinates of the GRCh37 PARs are given here:

`ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/par.txt`

3.6.2 Low-coverage and exome alignment and BAM processing

Authors: Shane A. McCarthy, Sendu Bala

Illumina low coverage and exome sequencing data was aligned by the Vertebrate Resequencing Informatics group at the Sanger Institute. Alignments were updated after Phase 1 to used that latest software and input files available at the time production began in late 2011.

Run-level alignment and BAM improvement: Fastq files were regularly downloaded from the 1000 Genomes ftp site based on the 'sequence.index' file coordinated by the 1000 Genomes Data Coordination Centre (DCC). The final set of BAM files

used for analysis were based on 20130502.phase3.analysis.sequence.index which contained 2,535 samples with both low coverage and exome data from 59,824 sequencing runs that passed quality controls (3.7). The sequence index is available here:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.analysis.
sequence.index
```

Data was aligned with bwa v0.5.9[15] to the GRCh37 (hg19) decoy reference (see Section 3.6.1). The reference fasta file was first indexed:

```
bwa index -a bwtsw $ref_fa
```

Then, for each fastq file, a suffix-array index (sai) file was created

```
bwa aln -q 15 -f $sai_file $ref_fa $fq_file
```

Aligned SAM files[16] were created using using 'bwa sampe' or 'samse' for paired-end or unpaired reads respectively. For paired-end reads, the maximum insert size was set to be 3 times the expected insert size.

```
bwa sampe -a $max_insert_size -f $sam $ref_fa $sai_files $fq_files
bwa samse -f $sam_file $ref_fa $sai_file $fq_file
```

SAM was converted to BAM, name-sorted, mate information fixed, coordinate-sorted and the MD tag added:

```
samtools view -bSu $sam | \
samtools sort -n -o - samtools_nsort_tmp | \
samtools fixmate /dev/stdin /dev/stdout | \
samtools sort -o - samtools_csort_tmp | \
samtools fillmd -u - $ref_fa > $bam
```

As in Phase 1, run-level alignment BAMs are improved in various ways to help increase the quality and speed of subsequent SNP calling that may be carried out on them. Reads were locally realigned around known indels using GATK IndelRealigner.

```
java $jvm_args -jar GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R $ref_fa -o $intervals_file \
```

```
-known $known_indels_file(s)

java $jvm_args -jar GenomeAnalysisTK.jar \
-T IndelRealigner \
-R $ref_fa -I $bam_file -o $realigned_bam_file \
-targetIntervals $intervals_file \
-known $known_indels_file(s) \
-LOD 0.4 -model KNOWNS_ONLY -compress 0 --disable_bam_indexing
```

The set of known indels was updated since Phase 1 to the include both the Mills-Devine double-hit high-quality indel set and the 1000 Genomes Phase 1 indel set. These files used are available here:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_
resources/ALL.wgs.indels_mills_devine_hg19_leftAligned_collapsed_double_hit.
indels.sites.vcf.gz
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_
resources/ALL.wgs.low_coverage_vqsr.20101123.indels.sites.vcf.gz
```

Base quality scores were then recalibrated using GATK CountCovariates and TableRecalibration.

```
java $jvm_args -jar GenomeAnalysisTK.jar \
-T CountCovariates \
-R $ref_fa -I $realign_bam -recalFile recal_data.csv \
-knownSites $known_sites_file(s) -l INFO \
-cov ReadGroupCovariate -cov QualityScoreCovariate \
-cov CycleCovariate -cov DinucCovariate \
-L '1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;X;Y;MT'

java $jvm_args -jar GenomeAnalysisTK.jar \
-T TableRecalibration \
-R $ref_fa -recalFile recal_data.csv \
-I $realign_bam -o $recal_bam \
-l INFO -compress 0 --disable_bam_indexing
```

The set of known sites for recalibration was updated since Phase 1 to dbSNP135, which includes sites from Phase 1.

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_
resources/ALL.wgs.dbsnp.build135.snps.sites.vcf.gz
```

Recalibrated BAMs were then passed through samtools calmd to fix NM tags and

introduce BQ tags which can be used during SNP calling[17].

```
samtools calmd -Erb $recal_bam $ref_fa > $bq_bam
```

Release BAM file production: The improved BAMs were merged together to create the release BAM files available for download. Release BAM files therefore contain reads from multiple readgroups.

Run-level BAMs have extraneous tags (OQ, XM, XG, XO) stripped from them, to reduce total file size by around 30%. Tag-stripped BAMs from the same sample and library were merged with Picard MergeSamFiles.

```
java $jvm_args -jar MergeSamFiles.jar \
INPUT=$tag_strip_bam(s) OUTPUT=$library_bam \
VALIDATION_STRINGENCY=SILENT
```

PCR duplicates are marked in library BAMs using Picard MarkDuplicates.

```
java $jvm_args -jar MarkDuplicates.jar \
INPUT=$library_level_bam OUTPUT=$markdup_bam \
ASSUME_SORTED=TRUE METRICS_FILE=/dev/null \
VALIDATION_STRINGENCY=SILENT
```

Duplicate-marked library BAMs from the same sample were merged with Picard MergeSamFiles.

```
java $jvm_args -jar MergeSamFiles.jar \
INPUT=$markdup_bam(s) OUTPUT=$sample_bam \
VALIDATION_STRINGENCY=SILENT
```

Sample BAMs were split into mapped and unmapped BAMs for release.

3.6.3 High-coverage PCR-free alignment and BAM processing

Authors: Ryan Poplin, Mauricio Carneiro

The data processing for the whole genome 2×250 PCR-free validation data (see section 3.3) differs from the data processing that was applied to the main project's

low-coverage and exome sequencing data in three aspects: First, we used BWA-MEM[18] instead of BWA-ALN[15]. Second, we explicitly apply an adapter clipping procedure because the longer read length found in this data resulted in an increased rate of reading through the adapter sequence. Finally, there is no marking of PCR duplicates since the data is PCR-free. The entire processing pipeline is as follows below.

1. Alignment to GRCh37 (hg19) decoy reference (3.6.1) genome.

   ```
   bwa mem -p -M -t $ref_fa $fq_file
   ```

2. Adapter clipping.

   ```
   java -jar MarkIlluminaAdapters.jar INPUT=in.bam OUTPUT=out.bam
   PE=true ADAPTERS=DUAL_INDEXED M=out.bam.adapter_metrics
   ```

3. Indel realignment. See `http://gatkforums.broadinstitute.org/discussion/38/local-realignment-around-indels` for command lines.

4. Base Quality Score Recalibration. See `http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr` for command lines.

3.7 Quality control of project alignment files

Authors: Xiangqun Zheng-Bradley, Laura Clarke

The sequence data was aligned to GRCh37 as described in section 3.6.2. The Data Coordination Centre (DCC) undertook several quality control (QC) steps on the aligned BAMs to check for low quality data or sample mix-ups.

The consortium established minimum coverage criteria for samples to be included in the analysis. For low coverage data, the required coverage level was 3× non-duplicated, aligned coverage. 3 samples failed to meet this and were excluded from the analysis. The minimum criterion for exome data was that more than 70% of the target regions were covered by 20× or greater of sequence reads. This was calculated using the Picard tool CalculateHsMetrics. 16 samples did not meet the criteria and were excluded. These statistics are available on the FTP site here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/alignment_indices/20130502.exome.alignment.index.HsMetrics.gz`

We have observed that samples with unbalanced ratio of short insertion and deletion tend to be indicative of low quality sequence data. A basic algorithm was used to calculate this ratio on all alignments. The process highlighted 3 low coverage samples and 6 exome samples with an unusually high ratio of insertions to deletions (greater than 5). These samples were excluded from our analyses. Both the statistics for each sample and the code used to calculate this ratio can be found at:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/alignment_indices/exome_indel_ratio_check.20130502.txt`
`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/alignment_indices/indel_ratio_check.20130502.txt`
`https://github.com/lh3/samtools-legacy/blob/master/examples/chk_indel.c`

Sample contamination and sample mix-ups were evaluated using VerifyBamID[19]. VerifyBamID takes a subset of reads from a sample and compares them to known genotype data for the sample to calculate how likely the sample is to be contaminated with DNA from unintended source or represents a sample swap. VerifyBamID can also robustly identify sample contamination using a chip-free mode if no known genotypes are available for a given sample. Almost all samples had genotype data either from the Illumina OMNI 2.5M genotyping chip or the Affymetrix 6.0 or both and are available here:

`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/`

For the small number of samples where no known genotypes were available, the chip-free mode was used during the evaluation. When a sample was analysed using VerifyBamID, only genotype data from the corresponding population were used. The calculation was performed at the sequencing run level to achieve a finer resolution of contamination likelihood. Contamination level 3% and 3.5% was used as cutoffs for low coverage and exome samples respectively; using these cutoffs, 22 low coverage and 96 exome samples were excluded from further analysis.

SNP genotype concordance between an earlier callset and OMNI 2.5M genotype data suggested that sample NA20816 is an outlier with poor concordance for all possible genotypes.

In order to ensure consistency across our analysed samples we only considered samples for which both low coverage and exome sequence data passed all our QC. 39 samples did not have both low coverage and exome sequence data so were excluded from further analysis.

After all QA steps, 2,535 samples passed and proceeded to the variant calling process.

4 Variant calling

The Phase 1 callset, in an effort to maintain a false discovery rate <5%, contained only high-quality biallelic SNPs, indels and large deletions. With method advancements in Phase 3, the 1000 Genomes catalogue now includes multi-allelic SNPs and indels, MNPs, complex substitutions, and a range additional structural variants, while maintaining the FDR<5% goal. Multiple callsets were integrated to provide a single set of haplotypes from the the Project. Details of the individual input callsets, some of which leverage assembly (either local or global) and haplotype aware calling, are given below, followed by details of the integration process in Section 5.

All the individual callsets can be found here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets`

4.1 Short variants – SNPs, indels, MNPs, complex substitutions

4.1.1 Baylor College of Medicine HGSC – SNPtools & Atlas

Authors: Zhuoyi Huang, Fuli Yu

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/bcm/`

We used SNPTools[20] for SNP calling in the low coverage whole genome sequencing and Atlas2[21,22] for SNP calling in the high coverage exome sequencing, and integrated both to produce a final whole genome SNP call set.

SNPTools is a suite of tools for high quality discovery, genotyping and phasing of SNPs in low coverage population sequencing. To perform a joint SNP calling across all samples, SNPTools first effectively reduces the alignment data of each sample and calculates the Effective Base Depth (EBD), integrating the mapping and base quality at each site. The reduced BAM file, or EBD file, contains essential information for SNP calling but only with about 1/16 of the BAM size. SNPTools then aggregates EBD files of all samples and discovers the polymorphic loci by applying the Variance Ratio Statistics, which scores the significance of the variant within the population. SNPs with high variance ratio score are included in the low coverage SNP call set. In order to increase the coverage in the whole genome region, we used both the

whole genome low coverage alignment data, and the off-target regions in the exome alignment data in the SNP calling.

The Phase 3 low coverage sequencing data includes 2,535 samples, and the alignment data is of size 80 TB, stored in the Cloud using Amazon Simple Storage Service (S3). The large sample size and data volume imposes a challenge for efficient and accurate SNP calling. We deployed and fully parallelised the SNPTools pipeline in the Cloud using Amazon Web Service, and achieved high efficiency by exploiting high performance Elastic Compute Cloud (EC2) instances and MapReduce algorithm. The entire low coverage SNP calling in the Cloud required two weeks of computing.

Atlas2 is a variant analysis pipeline optimised for variant discovery for high coverage data sequenced using different platforms. It employs logistic regression models, validated using whole exome capture sequencing data, to call SNPs and INDELs separately with high sensitivity and accurate genotypes. To call SNPs in Phase 3 high coverage exome sequencing data, we first performed the calling on each sample individually, and generated a master union of sites containing the SNPs from all samples. We then repeated the sample level calling but only at sites in the master union. The sample level call sets were then merged to produce the exome sequencing call set. Finally, the exome and low coverage call sets were integrated as a whole genome SNP call set.

We identified 72,956,744 SNPs from 2,535 Phase 3 genomes, and the average Ts/Tv ratio is 1.98. To compare with Phase 1 SNP call sets, we subset the call set with 1,047 samples shared in both Phase 1 and Phase 3. 55.61% of the SNPs called the Phase 3 call set are novel to Phase 1, and we rediscovered 89.81% of the SNPs in Phase 1. 91.30% of the missing SNPs are from samples with sequencing platform change and/or sequencing data change between Phase 1 and Phase 3.

4.1.2 Boston College – Freebayes

Author: Erik Garrison

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/bc/`

FreeBayes v0.9.9.2-6-gfbf46fc[23] was used to call variants from the refined alignments produced by standard process used in the project (Section 3.6.2). A number of parameters were set so as to improve the runtime of the process and simplify filtering of the intermediate results. Three observations of a candidate allele were required to be found in a single individual for the allele to be considered

(`--min-alternate-count 3`). Similarly, the quality sum of an alternate was required to be at least 50 (`--min-alternate-qsum 50`). Reads with mapping quality of 0 were excluded (`--min-mapping-quality 1`). And alleles with qualities less than 3 were excluded (`--min-base-quality 3`). Runtime was improved in the large sample set by limiting the depth of posterior integration (`--genotyping-max-iterations 10`). Additionally, contamination estimates from VerifyBamID (`http://genome.sph.umich.edu/wiki/VerifyBamID`) were supplied to improve genotype estimation via the `--contimation-estimates` flag. The contamination estimates used can be found here:

`https://github.com/ekg/1000G-integration/blob/master/resources/p3.exome_lowcov.per_RG.het_and_contam.contaminations`

The process was run over 13,621 genomic regions of approximately equal alignment content, merged, and then filtered with a simple quality filter requiring that the site QUAL was greater than 1.

4.1.3 Broad Institute – Unified Genotyper

Authors: Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Guillermo del Angel, Eric Banks

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/bi/*mapping*`

The Broad Institute produced a SNP and indel callset for both the low-coverage and exome samples using the GATK's UnifiedGenotyper in an identical manner as was employed in previous iterations of the project. This callset was included here as a baseline by which to compare the methodological advances of the project. See the Phase 1 paper[4] for a complete description of this method.

4.1.4 Broad Institute – Haplotype Caller

Authors: Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Guillermo del Angel, Eric Banks

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/bi/*assembly*`

(0) Defining the ActiveRegions (AR): Active regions are determined by calculating a profile function that characterises interesting regions likely to contain

variants.

A raw profile is first calculated locus by locus. The profile value is calculated as the probability that the position contains a variant as calculated using the reference-confidence model applied to the original alignment.

This operation gives us a single raw value for each position on the genome (or within the analysis intervals requested by the user).

However the final profile is calculated by smoothing this initial raw profile following three steps. The first two steps consist in spreading individual position raw profile values to contiguous bases. As a result each position will have more than one raw profile value that are added up in the third and last step to obtain a final unique and smoothed value per position.

1. First, a position profile value will be copied over to adjacent regions if enough high quality soft-clipped bases immediately precede or follow that position in the original alignment. Currently high-quality soft-clipped bases are those with quality score of Q29 or more. We consider that there are enough of such a soft-clips when the average number of high quality bases per soft-clip is 7 or more. In this case the site profile value is copied to all bases within a radius of that position as large as the average soft-clip length without exceeding a maximum of 50 bp.
2. Second, each profile value (including those generated in step 1.) is divided and spread out using a Gaussian kernel covering up to 50 bp radius centred at its current position with a standard deviation of 17.0 base pairs
3. Finally, each position final smoothed value is calculated as the sum of all its profile values after steps 1 and 2.

Then the resulting profile line is cut in regions where it crosses the non-active to active threshold currently set to 0.002. Then we make some adjustments to these boundaries so that those regions that are to be considered active, with a profile running over that threshold, fall within the minimum (fixed to 50 bp) and maximum region size:

1. If the region size falls within the limits we leave it untouched,
2. Or if the region size is shorter than the minimum it is greedily extended forward ignoring that cut point and we come back to step 1. Only if this is not possible because we hit a hard-limit (end of the chromosome or requested analysis interval) we will accept the small region as it is.

3. Or if it is too long we find the lowest local minimum between the maximum and minimum region size. A local minimum is a profile value preceded by a large one right up-stream (-1 bp) and an equal or larger value down-stream (+1 bp). In case of a tie, the one further downstream takes preference. If there is no local minimum we simply force the cut so that the region has the he maximum active region size.

Of the resulting regions, those with a profile that runs over this threshold are considered active regions and progress to variant discovery and or calling whereas regions whose profile runs under the threshold are considered inactive regions and are discarded.

(1) Finalise the Active Region: Before we start to do any work with AR we do some clean up with its reads:

1. Remove bases at each end of the read (hard-clipping) until there a base with a call quality equal or greater than minimum base quality score (user parameter -mbq, 10 by default).
2. Include or exclude remaining soft-clipped ends. Soft clipped ends will be used for assembly and calling if and only if the user has not requested their exclusions (using -dontUseSoftClippedBases) and the read and its mate map to the same chromosome and they are in the right standard orientation (i.e. LR and RL).
3. Clip off adaptor sequences of the read if present.
4. Discard all reads that after 1-3 the original alignment does not overlap with the AR anymore.
5. Downsample remaining reads to a maximum of 1000 reads per sample but respecting a minimum of 5 read start per position. The user has no control on this procedure and is performed after any downsampling by the traversal itself (-dt, -dfrac, -dcov etc.)

(2) We determine haplotypes and list of potential variant sites: A haplotype is a possible reconstruction of the AR based on the input read data. Each AR can have several haplotypes due to:

1. real diversity on polyploidy (including CNV) or multi-sample data,
2. possible allele combinations between non-totally linked variants sites within the AR,

3. or sequencing and mapping errors.

In order to generate a list of possible reconstructions we build an assembly graph for that AR and we select the best (seemly most likely) paths across that graph. The type of graph that HC generates is a concept called the read threading graph (RTG). RTG is strongly based on the De Bruijn graph that is common place in genome assembly methodology.

RTG reconstruction's main parameter is the k-mer size k, where a k-mer is a continuous sequence of k bases. The user can indicate several k-mer sizes (argument -kmerSize , 10 and 25 by default) and HC will attempt to compose a RTG for each of those k. The final set of haplotypes will be the union of the sets obtained using each k, following the procedure below:

1. Construct the assembly graph using the reference sequence which covers the active region and all read base sequences. The resulting directed graph will have one node each presenting a base and connecting edges indicating the sequential relationship between the adjacent bases; for every edge the source vertex' base of an edges directly precedes the target vertex' base in at least one haplotype. Edges are labelled with the number of reads that support the existence of that sequential relationship.

2. Removal of noise due to errors: we prune sections of the graph with low support presumably caused by stochastic errors. This can be controlled by with the -minPruning INT argument. For example, if minPruning == 3, linear chains in the graph (linear sequence of vertices and edges without any branching) where all its edges have less than three reads supporting its existence are removed. Pruning won't act on a segment if there is a few samples with support larger than the one indicated with minPruning. The minimum number of supporting samples is controlled also by the user with -minPruningSamples INT (1 by default).

3. Recovery of dangling tails and heads due to limitations graph assembly step 1 or sequencing errors at the ends of reads.

4. Select best haplotypes: in order to put a limit to the amount of computation needed there is a limit to the number of haplotypes that will be considered per each k for further analyses. This is controlled by the user through the argument -maxNumHaplotypesInPopulation INT (set generously to 128 by default). If there is the need to discard some we only use the best ones based on their score or likelihood. This score is calculated as the product of transition probabilities

of its edges where the transition probability of an edge is in turn computed as the number of reads supporting that edge divided by the sum of the support of all edges that share that same source vertex.

If none of the k-mer size provided results in a viable graph (complex enough and without cycles we try the operation with larger k-mer sizes. More concretely we get the largest k provided by the user and add 10 to it up to 6 times until we get a viable assembly graph. A graph is said to be complex if the number of non-unique k-mers is less that 4-fold the number of unique k-mers found in the data.

Once we have a list of plausible haplotypes we determine the variation sites using its CIGAR string that it turn is reconstructed by applying Smith-Waterman alignment (SWA) of the each haplotype sequence with the reference sequence across the ER. In the case of indels of repeated units their start location will always be the left-most (lowest) possible.

(3) Active region trimming: In order to save computational time, we reduce the size of the active region to include only variation observed in the haplotypes with some padding around it.

We then filter out bad reads: we filter reads that won't be further considered for genotyping but they are ok to be used for assembly. For a read to be considered for genotyping it requires all of the following:

- length after being trimmed at the AR borders (using the original alignment) must be greater or equal to 10 bases.
- mapping quality is at least Q20.
- its mate must be mapped on the same chromosome.

(4) Calculate per-read likelihoods: We use a PairHMM model to calculate the likelihood of each kept read versus each candidate haplotype[24]. This step is identical to the UnifiedGenotyper

(5) Calculate per genotype likelihoods: For each sample and candidate variation site discovered in step 2) or in the input variant file if working in GGA mode, we calculate the likelihood of each read vs each allele at that site. We marginalise the likelihood of a read given a variant allele as the maximum likelihood of that read on any of the haplotypes that contain that allele.

Then we use the standard UG formulation to calculate the genotype likelihoods and assign the most likely genotype. Please refer to the appropriate UG/HC presentation for details.

We use the approach described by Li[25] to calculate non-reference allele count posterior probabilities (Methods 2.3.5 and 2.3.6) but extended to handle multi-allelic variation. Then, we estimate the confidence on whether the variant exists at that position (QUAL column in VCF output) as the Phred scale of the posterior probability of the non-reference allele count to be exactly 0. The most likely estimate for the non-reference allele count is also reported and the corresponding frequency (MLEAC and MLEAF).

Finally variant call is annotated with requested info- and genotype fields that form part of the final VCF output.

(6) Post-calling filtration of variants: In the same manner as with the UnifiedGenotyper approach we employed the Variant Quality Score Recalibrator[26] to filter the SNP and Indel variants identified by the HaplotypeCaller.

4.1.5 University of Michigan – GotCloud

Authors: Hyun Min Kang, Adrian Tan, Goo Jun, Mary Kate Wing, and Gonçalo R. Abecasis

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/um/`

The GotCloud SNP and Indel calling pipeline was used to produce variant calls contributed by University of Michigan. The SNP calling pipeline first computes genotype likelihoods for each low-coverage and deep exome BAM files using the default samtools genotype likelihood model, after adjusting by per-base alignment quality (BAQ) and trimming a lower-quality end of paired-end reads. Genotype likelihoods are merged across low-coverage and exome sequence reads. 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 to obtain maximum likelihood estimates of allele frequency at each locus. We compared likelihoods with no variant to the likelihood with 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-scale quality score of 5) with neutral allele frequency spectrum under a constant population site at average heterozygosity between two chromosomes to be

1 in 1,000 bp.

The candidate variant sites identified from the initial discovery step were filtered based on multiple features reflecting site-specific sequencing quality, such as sequencing depth, and the fraction of bases with low quality scores, and features reflecting quality of the evidence for a variant, such as the fraction of bases with reference allele in heterozygous samples (allele balance) and the correlation between observed alleles and the read direction (strand bias). GotCloud uses Support Vector Machine (SVM) with radial basis function (RBF) kernel to train the classifier distinguishing likely true variants from likely false variants. A list of positive and negative examples from the union of array-based polymorphic sites identified from the HapMap project[27] and Omni2.5 SNP array. Lists of likely false positives are seeded with sites that fail multiple stringent hard-filters.

The Indel calling pipeline first identifies variant sites observed more than once from aligned sequences and the variant sites are merged across the samples. Genotyping is performed from each candidate sites across all samples, by aligning the sequence reads with each possible allele using a pair Hidden Markov Model (HMM). SVM filtering similar to SNP filtering is applied, and variants around the k-mers showing excessive heterozygosity in male chromosome X were additionally filtered out. Our SNP and Indel calling procedures are implemented in the GotCloud variant calling pipeline available at `http://www.gotcloud.org`.

4.1.6 Oxford University – Platypus

Authors: Andrew Rimmer, Hang Phan, Gerton Lunter

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/ox/`

Platypus[28] is a haplotype-based variant caller. 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 from 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.

Several filtering steps improve the robustness of calls and reduce the number of

spurious candidates. To remove poorly or ambiguously mapped reads, Platypus requires a minimum mapping quality of 20 on the Phred scale. To reduce the impact of non-independent errors, duplicate reads are removed.

Platypus considers variant candidates 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 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[29]. 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 5). Additional filters are used to remove variants which are only supported by low-quality reads, or reads on the forward or reverse strand.

To call variants from the data presented in this paper, we applied Platypus with default parameters, except for the buffer size which was reduced to 1000 to cope with the large number of input BAM files.

4.1.7 Oxford University – Cortex

Authors: Zamin Iqbal, Xiangqun Zheng-Bradley, Chunlin Xiao, Anthony Marcketta, Adam Auton

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/cortex/`

Cortex is a *de novo* De Bruijn graph assembler that allows simultaneous assembly of multiple samples and variants to be called without reliance on mapping of reads to

a reference genome. Calls for Phase 3 of the 1000 Genomes Project were generated using Cortex v1.0.5.15[30], which is archived here:

`http://sourceforge.net/projects/cortexassembler/files/cortex_var/previous_releases/CORTEX_release_v1.0.5.15.tgz`

Some key resource files are available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120814_cortex_resources/`

The pipeline was as follows, and all scripts have been saved in the 'scripts/1000genomes' directory of the Cortex release described above. A detailed description of the pipeline used to create the Cortex callset can be found here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/cortex/README_cortex_pipeline_20150213`

4.1.8 Sanger Institute – SAMtools/BCFtools

Authors: Petr Daněček, Shane A. McCarthy

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/si/*samtools*`

Raw calls were produced with samtools (version 0.1.19-40-g06aeb0c) and bcftools (version 0.1.19-47-gfec2433)[25]. Samtools was used to generate all-site all-sample BCF files (samtools mpileup -EDVS -C50 -pm3 -F0.2 -d 10000) from all 2,535 low-coverage sample BAM files. Bcftools was subsequently used to call variants (bcftools view -Ngvm0.99). On chromosome X, male samples were treated as diploid in the pseudo-autosomal regions (X:60001-2699520 and X:154931044-155270560) and haploid otherwise using the '-s' option in bcftools view. Calls were then filtered using 'bcftools som' with default parameters (bcftools-c3d530e/htslib-e91d10b). The SOM (Self-Organising Map) model was trained on 1000G Phase 1 calls using the annotations HWE (Hardy-Weinberg equilibrium p-value), MQ0 (Number of reads with zero mapping quality), PV2 (Mapping-quality bias p-value), QBD (quality-by-depth), RPB (read position bias), VDB (variant distance bias) and the cutoff was chosen at 99.5% sensitivity.

4.1.9 Sanger Institute – SGA-Dindel

Authors: Jared Simpson, Kees Albers, Shane A. McCarthy, Richard Durbin

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/si/*sga-dindel*`

SGA: The SGA variant calling algorithm[31,32] begins by finding k-mers present at least 30 times, and no more than 10,000 times, in the sequence reads that are not present in the reference genome. These k-mers are used to seed an assembly and re-alignment process that has three stages. First, the de Bruijn graph local to the seed k-mer is traversed to assemble candidate variant haplotypes. Second, the candidate haplotypes are mapped to the reference genome. Third, reads matching a candidate haplotype are extracted from the FM-index and the set of candidate haplotypes, their reference mappings and the extracted reads are passed to the DINDEL framework for evaluation. These three steps are described in more detail below. Within each description, we include a link to source code for the version of the software that was used to generate the variant calls.

Haplotype Assembly: Haplotype assembly begins by adding the seed k-mer to the de Bruijn graph and adding the k-mer to an exploration queue. The algorithm then processes the exploration queue until it is empty or pre-defined computational limits are reached. At each step, the k-mer at the front of the queue is processed. The suffix or prefix neighbours of the k-mer are found by querying the FM-index using the procedure described in[33]. Neighbour k-mers seen at least 10 times in the reads are added to the graph. Each k-mer added to the graph is checked for whether it also appears in the reference genome. If the k-mer appears in the reference it is annotated as an "upstream" or "downstream" junction vertex, depending on the direction of graph exploration. The non-junction k-mers are enqueued for further exploration of the graph.

Once the graph exploration is complete, paths through the graph from every upstream k-mer to every downstream k-mer are found. The strings corresponding to these paths form the initial candidate haplotype set. The complete source code for the haplotype assembly algorithm can be found here:

`https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/DeBruijnHaplotypeBuilder.cpp#L42`

Next we perform quality checks on the haplotypes. Each haplotype must contain at least $k/2$ consecutive non-reference k-mers. This check identifies short, low-quality haplotypes.

Additionally, we calculate the maximum value of k for which the haplotype sequence forms a complete, unbroken path through the k-de Bruijn graph of the sequence reads or the reference genome. Call these values maxreadk and maxrefk respectively. We discard a haplotype if maxrefk $\geq$ 41 or maxrefk $\geq$ maxreadk or maxreadk $-$ maxrefk $<$ 10. This check ensures that the haplotypes have stronger support in the graph of the sequence reads than the graph of the reference genome. If any haplotype fails a quality check the entire set is discarded and no variants are output. The source code for this function is here:

`https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/GraphCompare.cpp#L288`

Next, an attempt to assemble the reference sequence that corresponds to each candidate haplotype. To do this, we perform a directed walk through the reference de Bruijn graph from the first k-mer to the last k-mer of every candidate haplotype. Any sequences that successfully assemble are added to the candidate haplotype set. The source code for this function is here:

`https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/GraphCompare.cpp#L402`

Haplotype Alignment: Each candidate haplotype is aligned to the reference genome using a seed-and-extend approach. 31-mers are used to seed the alignment. If a haplotype 31-mer appears in the reference no more than 4 times, it triggers an alignment between the haplotype and the reference. The alignment uses standard dynamic programming with an affine gap penalty. To accept an alignment it must have at least 50 aligned bases, at least 95% sequence identity and no more than 8 variation events (all consecutive substitutions, insertions or deletions are considered to be a single event). If more than 10 alignments are found, the haplotype set is discarded. The reference sequence for each alignment location is added to the candidate haplotype set. The source code for this function is here:

`https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/HapgenUtil.cpp#L65`

Read Extraction: Next, we extract reads from the FM-index that share a 31-mer with any haplotype in the candidate set. As some haplotypes contain highly repetitive 31-mers we discard 31-mers that occur more than 5000 times in the reads. Additionally, if more than 100,000 reads are extracted the process stops and the haplotype set is discarded. The extracted reads, along with the candidate haplotypes, are input into the DINDEL algorithm described below. The source code for this function is here:

https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/HapgenUtil.cpp#L388

DINDEL: The DINDEL variant calling algorithm largely follows the framework described in[29]. The DINDEL algorithm consists of the following stages to call variants for a given seed k-mer and associated set of candidate haplotypes. First, each read is aligned to each candidate haplotype using a modified version of the original implementation of DINDEL as described below, yielding a likelihood $P(r_i|H_j)$ for each read i and haplotype j. Second, a model selection procedure is used to identify the set of candidate haplotypes that have sufficient support from the reads

https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/DindelRealignWindow.cpp#L3924

At each step of the model selection procedure, the improvement to the log-likelihood achieved by adding a new candidate haplotype to the model is recorded and is used to calculate the quality score of the variants represented by that candidate haplotype. Next, assuming diploid individuals and explicitly taking into account which individual each read is observed for, an expectation-maximization algorithm is used to estimate the haplotype frequencies for the subset of candidate haplotypes that survives the model selection procedure. In the last step, the selected candidate haplotypes and corresponding haplotype frequency estimates are used to produce variant calls

https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/DindelRealignWindow.cpp#L1148

The procedure allows for the possibility that a selected candidate haplotype may align to multiple positions in the reference genome with different probabilities. Furthermore, for a given reference alignment, multiple selected haplotypes may provide support the same variant. Thus, this approach may call variants at multiple positions in the reference genome from the support for a single selected candidate haplotype.

Modified DINDEL read-haplotype realignment model
The original DINDEL realignment model[29] was modified for improved efficiency.

https://github.com/jts/sga/blob/629e0875d8efcbcdf4d3a57375d239908c7c619c/src/GraphDiff/DindelHMM.cpp

Instead of considering all possible alignments of the read to a candidate haplotype as done originally, now only 3 candidate alignments are considered. To determine these candidate alignments, first a hash of 8-mers to position for both the haplotype and the read are constructed. Next, the three relative positions of read and haplotype

for which the number of matching hashes is largest are considered for probabilistic realignment by the HMM. Second, the relative positions of read and haplotype are considered to deviate no more than four nucleotides from the three relative positions of read and haplotype. Furthermore, the largest sequencing error indel is assumed to be four nucleotides long; the alignment log-likelihood is bounded from below to prevent a single poorly matching read from disproportionately influencing the inference. Finally, the relative position of read and haplotype with the highest probability is reported as the alignment probability $P(\text{Read sequence} = r|\text{Haplotype sequence} = h)$.

Pipeline and filtering: The above method was used on the Phase 3 low-coverage data using SGA version 0.10.8 using the pipeline descibed here. The 2,535 low-coverage mapped BAM files were split into 5 Mbp chunks overlapping by 1 kbp and converted to fastq. Reads shorter than 75 bp were removed using

```
sga preprocess --min-length 75
```

For each 5 Mbp chunk of the genome, all fastq files from all 2,535 samples were merged into a single fastq with a corresponding 'popidx' index file to track which reads within the merged fastq belonged to which sample. The merged fastq files were then indexed using

```
sga index --no-reverse -a ropebwt -t 8
```

A copy of the reference genome was prepared by randomly replacing ambiguous bases by A, C, G or T and creating a sampled-suffix-array with the following commands:

```
sga preprocess --permute-ambiguous $ref > $ref.permute.fa
sga gen-ssa $ref.permute.fa
```

For each 5 Mbp chunk, variants were called using

```
sga graph-diff --debruijn --low-coverage -k 61 -x 30 -m 10 -t 16 \
--variant $fq --reference $ref.permute.fa -p $chunk
```

To aid filtering, a diploid segregation model was then applied to the calls using

```
sga haplotype-filter -t 4 -o $chunk.filt.vcf --reads $fq \
--reference $ref.permute.fa $chunk.strings.fa $chunk.calls.vcf
```

The chunks were coordinate sorted and concatenated together with duplicate sites removed. Sites where the reference allele differed from the original reference due to the permuting of ambiguous bases were removed.

GATK (v2.4-9-g532efad) Variant Quality Score Recalibration (VQSR)[26] was used to filter SNPs and indels. Training was based on the annotations LM (Log-likelihood ratio statistic using diploid segregation model), O (Number of reads used in segregation test), SB (Strand bias), VarQual (Variant quality), VarDP (Number of reads containing the variant). For SNPs, 1000G Omni2.5 and HapMap 3.3 sites were used for training. A truth sensitivity cutoff of 93% was chosen for SNPs. For indels, the Mills-Devine 1000G gold standard sites were used as truth, training and known. A truth sensitivity cutoff of 50% was chosen for indels.

Calls were not produced for 3 problematic 5 Mbp chunks (1:139972001-144972000, 9:64987001-69987000 and 16:44991001-49991000). These chunks are in regions of the genome where there is a long reference gap and many DGV structural variants.

4.1.10 Stanford University – Real Time Genomics

Authors: Suyash S. Shringarpure, Andrew Carroll, Francisco De La Vega, Carlos D. Bustamante

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/stn/`

SNP calling was performed on the low-coverage BAM files from the project. We used the rtgVariant population caller (version 3.1.2, Real Time Genomics, San Bruno, CA). rtgVariant uses a Bayesian framework for variant calling, including a haplotype-aware method for complex calls (small indels and MNPs), and recalculates site specific priors based on allele counts on the population sample until convergence, assuming Hardy-Weinberg Equilibrium (HWE)[34]. We grouped the samples by population for calling to avoid strong violations in HWE. rtgVariant also preforms pedigree-aware calling simultaneously with unrelated data[34], and we thus provided pedigree information for the trios in the samples for variant calling. To maximise parallelisation, the BAMs were split by population and chromosome – variant calling was performed in 572 parallel jobs (26 populations $\times$ 22 autosomes) and the final callset was obtained by merging VCFs from all jobs. The DNAnexus platform was used as an interface to the Amazon EC2 cloud to perform cloud-based variant calling using the BAMs stored in Amazon S3 (`http://aws.amazon.com/1000genomes/`). More details about about our analysis, including cost, runtime, scalability etc. can be found in Shringarpure *et al.*[35].

Variant calls were filtered using the rtgVariant Adaptive Variant Rescoring method (which uses a random forest algorithm to evaluate a number of covariates and provides a probability that a call is correct[34]) with an empirically defined cutoff of 0.05 to obtain a variant set with good quality metrics. We discovered 57 million SNPs in the 2,535 samples. The Ts/Tv ratio for these sites was 2.06 The false-positive rate on these sites, computed using the OMNI genotypes was 2.5%. 95.8% of HapMap sites and 76.5% of the 1000 Genomes Phase 1 variant sites were rediscovered in our call set.

4.2 Micro-satellites (STRs)

Two callsets were made to call microsatellite or short tandem repeat (STR) events. Neither callset was included in the final integrated callset, however the lobSTR callset was included during the integration process (Section 5), but removed at the filtering stage due allelic dropout (discussed in the lobSTR section below) causing too high a false discovery rate for the project goals.

4.2.1 LobSTR

Authors: Thomas Willems, Melissa Gymrek, Yaniv Erlich

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/strs/full/ALL.wgs.v3_lobSTR.20130502.microsat.integrated.genotypes.vcf.gz`

Call set generation

We generated BAM files containing STR-spanning reads by running GitHub version 8a6aeb9 of the lobSTR[36] genotyper on all FASTQ files listed in the 20130502.phase3.sequence.index file. The genotyper was run using the options `fft-window-size=16`, `fft-window-step=4`, and `bwaq=15` to align the reads to a custom reference of ∼700,000 STR loci generated previously[37]. We then used SAMtools[16] to remove duplicates from each individual BAM file before merging them by population. Finally, we ran v3.0.3 of the lobSTR allelotyper on all population BAM files concurrently to generate the STR genotypes. In addition to using the normdup option and v2.0.3 of the Illumina stutter model, we mitigated the effect of poorly aligned reads using the filters `min-read-end-match=5`, `min-bp-before-indel=7` and `maximal-end-match=15`. The resulting call set contains STR genotypes for 2,588 individuals and over 670,000 loci, with an average of over 478,000 calls per sample.

Quality assessment

To assess the quality of the call set, we compared the length of the lobSTR genotypes to those obtained using capillary electrophoresis for 176 of the Phase 3 samples[38]. These comparisons involved 281 annotated STRs that comprise the Marshfield microsatellite panel and are some of the most polymorphic STRs. To map capillary lengths to the associated lobSTR length, we first computed offsets as described previously[37]. For homozygous capillary genotypes, the lobSTR calls were correct in nearly 92% of cases (see table below).

Call	X/X	Y/Y	X/Y	Y/Z
Capillary genotype = X/X	8121 (91.81%)	356 (4.03%)	355 (4.01%)	13 (0.15%)

In contrast, for heterozygous capillary genotypes, lobSTR calls were only correct in 68.5% of the cases (see table below).

Call	X/X or Y/Y	Z/Z	W/Z	X/Z or Y/Z	X/Y
Capillary genotype = X/Y	14732 (68.58%)	948 (4.41%)	75 (0.35%)	713 (3.32%)	5013 (23.34%)

Incorrect calls for these heterozygous loci primarily stemmed from an issue known as allelic dropout, namely that a genotype X/Y was called as either X/X or Y/Y. However, this issue is largely expected as the bulk of the STR calls are solely based on a few reads. The low coverage also complicates distinguishing PCR stutter errors from true alleles, an issue that likely accounts for a substantial fraction of the remaining genotyping errors.

As an additional quality metric, we assessed the concordance of calls for 5 trios in the dataset with Mendelian inheritance. Because of their unusually high coverage relative to the other samples, we omitted the two high coverage trios and solely used the following samples for the assessment: HG00702, HG00656, HG00657 (trio 1), HG02024, HG02026, HG02025 (trio 2), NA19675, NA19679, NA19678 (trio 3), NA19685, NA19661, NA19660 (trio 4), HG00733, HG00731 and HG00732 (trio 5). Overall, nearly 93% of the calls were concordant with Mendelian inheritance, a figure that increased with coverage and plateaued at roughly 97%. The higher concordance obtained from the trio comparison likely stems from the fact that the majority of STRs in the reference are substantially shorter and less polymorphic than those used in the capillary comparison. As a result, the impact of allelic dropout at most loci in the call set is largely mitigated.

Filtering and applications

We provided the call set in an unfiltered fashion to maximise its potential utility. However, users should be aware of its inherent limitations. In particular, as demonstrated by the capillary data comparison, the individual genotypes in the call set frequently suffer from allelic dropout. As a result, for long or highly polymorphic loci, it may be useful to further filter by coverage using the DP FORMAT field contained in the VCF to ensure that both alleles were likely sampled. Nonetheless, the call set is particularly useful for obtaining statistics about each locus' frequency spectrum, degree of population differentiation and variability. In the absence of large length differences between alleles, the impact of allelic dropout on these characteristics is minimal, as dropout should occur fairly evenly for each allele.

The lobSTR callset was included during the integration process, but not included in the final integrated release.

4.2.2 RepeatSeq

Author: David Mittelman

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/strs/full/ALL.wgs.bwa_repeatseq.20130502.microsat.exome.genotype.vcf.gz`

RepeatSeq is an open source microsatellite variant caller (`https://github.com/adaptivegenome/repeatseq`) for genotyping 1-6-mer microsatellite repeats from Illumina resequencing data. Using flanking reads around specified repeat loci in a BAM file, RepeatSeq's genotyping model incorporates an experimentally derived error profile that considers repeat tract length, unit length, and base quality. For repeat loci in the human reference, we previously identified 700,000 sites using Tandem Repeat Finder and a second-order Markov framework outlined in Willems *et al.*[37].

Callsets were made for both the exome and low-coverage data from 2,535 individuals, including the 31 related individuals which have been removed from the main integrated callsets. Microsatellites genotyped are between 1-6 bp repeats, as per the functionality of RepeatSeq. In addition to the default command line, the parameter "-emitconfidentsites" was used to generate reference allele output in the VCF where genotypes were possible.

The RepeatSeq callset was not included in the integration process and is not in the final integrated release.

4.3 Structural variants (SVs)

Structural variant (SV) discovery and genotyping was performed on Phase 3 samples using a combination of 9 algorithms designed to identify deletions (DEL), duplications (DUP), multi-allelic copy number variants (mCNV), inversions (INV), mobile element insertions (MEI), and nuclear mitochondrial insertions (NUMT). The following sections provide an overview to their application. Further details on SV analysis methodologies can be found in the accompanying companion manuscript[39].

4.3.1 Breakdancer

SV Type(s): DEL
Authors: Wanding Zhou, Zechen Chong, Xian Fan, Klaudia Walter, Ken Chen

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/breakdancer/`

Deletion call-set generation: BreakDancerMax[40] (BD, v1.1.2) was run on all whole genome sequenced samples, and deletion calls were made by chromosome and separately for each population using reads with mapping quality greater or equal to 20. Insert size distributions were analysed for each library separately using a 1 Mb region on chr20 (chr20:10000000-11000000) to determine thresholds as upper cut-offs in the BD config files. The raw BD calls were filtered for deletion size (<50 bp and >1 Mb), for estimated read depth ratio (< 0.75), for number of spanning read pairs ($\geq$20), for regions around centromeres ($\pm$1 kb), for regions around assembly gaps ($\pm$50 bp) and for alpha satellite regions. The read depth (RD) ratio was calculated as the average RD of the samples that supported the deletion divided by the average RD of the samples that did not support the deletion. Deletions were then merged across all samples using 50% reciprocal overlaps and connected components and systematically genotyped using a software package called BreakDown (unpublished).

4.3.2 Delly

SV Type(s): DEL, DUP, INV
Authors: Tobias Rausch, Markus Fritz, Adrian Stütz, Jan Korbel

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/delly/`

Deletion and Duplication call-set generation: Delly[41] was run separately per

population on all Phase 3 low coverage WGS samples to identify deletions and duplications. All precise and imprecise Delly deletion predictions from the 26 populations were merged into a single SV site list using a 70% reciprocal overlap threshold and a maximum breakpoint offset of 250 bp. In each cluster, the paired-end mapping based call with the highest support was selected for the Delly's final candidate deletion sites list. Read-depth of all candidate deletions was annotated using 'cov', an auxiliary tool from the Delly package. The raw read-depth values were normalized for GC-content, mappability and median total coverage across samples and used to derive Gaussian Mixture Models for genotyping across the entire sample set.

Inversion call-set generation: Delly[41] was used separately for each population of the 1000 Genomes Project sample panel to identify inversions. Discovered population specific inversions sites were subsequently integrated into a merged inversion site list using a strict 90% reciprocal overlap criterion and a breakpoint offset smaller than 50 bp. The merged inversion site list was genotyped across the entire 1000 Genomes Project Phase 3 cohort using counts of inversion-supporting and reference-supporting read pairs.

4.3.3 Variation Hunter

SV Type(s): DEL
Authors: Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/variation_hunter/`

Deletion call-set generation: VariationHunter[42] deletion discovery considered all discordant mapping locations (paired-end reads exhibiting mapping spans more than 4 standard deviations above the inferred mean insert size) from mrFAST and BWA read alignments. To generate an initial callset we considered only those candidate sites with support of at least 2 read pairs, whereby we required an average edit distance of maximum 3 per read. We then applied several filters to reduce false positives: 1) we scaled the minimum read pair threshold for each sample according to the depth of coverage; 2) removed deletion calls overlapping segmental duplications >30% (reciprocal overlap criterion), 3) removed deletion calls that also show inverted duplication or inverted repeat insertion signals, and 4) required the read depth within the deletion interval to drop, consistent with the deletion event.

4.3.4 CNVnator

SV Type(s): DEL, DUP
Authors: Alexej Abyzov, Mark Gerstein

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/cnvnator/`

Deletion and Duplication call-set generation: Deletion and duplication calls with CNVnator[43] were made with standard parameters. Read depth (RD) signals were corrected for GC bias. Neighbouring read pairs showing abnormal read mapping with a mapping quality of at least 10 were used with the read depth calls to adjust breakpoints to reflect the more precise breakpoint inference of paired-end mapping when compared to read depth analysis. For each sample we subsequently selected confident CNVnator calls as follows: 1) calls having paired-end support; 2) calls with p-values less than 10^{-5} (to accounts for multiple hypothesis testing, i.e., calling in $\sim$2500 samples), and with $q0 < 0.5$; 3) deletion calls with p-values less than 10^{-5} and $\mathrm{rd} \times (1 + q0) < 0.75$, whereby rd is the read depth normalised to genome average, and $q0$ is fraction of reads mapped with 0 (zero) mapping quality. We merged CNV calls for individuals within each population. For CNVnator site merging we initially clustered confident overlapping calls and averaged coordinates of each bound, pursuing the merging initially by population and then across the entire sample set.

4.3.5 Read-Depth (dCGH)

SV Type(s): DEL, DUP, mCNV
Authors: Peter Sudmant, John Huddleston, Brad Nelson, Evan E. Eichler

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/uw/`

Deletion, Duplication, and mCNV call-set generation: The UW read-depth based call-set using digital comparative genomic hybrdization (dCGH) was generated based on remapping reads from all individual genomes with the mrsFAST read aligner[44]. Reads were first subdivided into their 36 bp non-overlapping constituents to normalise among the different read lengths represented in the 1000 Genomes Project dataset. After mapping, read-depths were quantified for each genome and recalibrated to take into account GC-associated coverage biases introduced by library construction[45]. Copy number was estimated in adjacent windows of 500 bp of unmasked sequence using a calibration curve based on regions of known copy number. Genomes were then assessed for overall quality using a number of QC metrics[45] with

a total of 2,169 samples passing all filters for analysis.

4.3.6 Genome STRiP

SV Type(s): DEL, DUP, mCNV
Authors: Bob Handsaker, Steve McCarroll

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/genome_strip/`

Deletion, Duplication, and mCNV call-set generation: Genome STRiP[46] (version 1.04.1225) was used to perform deletion discovery and genotyping. To increase sensitivity, deletion discovery was performed in five batches of approximately 500 samples each and a more stringent protocol for eliminating redundant polymorphic deletions was employed. For mCNVs and duplications, a novel read-depth pipeline based on Genome STRiP version 1.04.1375 was used that performed genotyping in 5kb overlapping windows across the genome followed by boundary refinement. Polymorphic sites ascertained from this pipeline that were confidently called as biallelic deletions were added to the Genome STRiP deletion call set if they had less than 50% reciprocal overlap with any site ascertained from the Genome STRiP deletion pipeline. For the other polymorphic CNV sites, copy-number genotype likelihoods were generated using Genome STRiP in a total of 2,356 samples passing quality control filters.

4.3.7 Pindel

SV Type(s): DEL
Authors: Kai Ye, Wubbo Lameijer, Klaudia Walter

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/pindel/`

Deletion call-set generation: Pindel[47] (version 0.2.5a2) was run across all Illumina paired-end samples in chunks of 300 kb with the following parameters: `-w 0.1 -x 5 -B 0 -T 4`. Regions around the centromeres were excluded. Split read based deletion calls appearing in at least 5 samples and with more than 5 reads from both strands were collected for downstream analysis, i.e. effectively removing SVs with allele counts of 1 to 4 to ensure high specificity (taking into account the lower specificity of split read based deletion discovery compared to other SV discovery modes).

Note, the Pindel callset was not included in the main integrated callset. However, it was included in the SV analyses including the companion paper[39] where is was run through the same MVNcall pipeline (Section 5.4.3) to phase onto the haplotype scaffold as other structural variant classes.

4.3.8 MELT

SV Type(s): MEI
Authors: Eugene J. Gardner, Scott E. Devine

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/umaryland/`

MEI call-set generation: Mobile element insertions (MEIs) were detected with the Mobile Element Locator Tool (MELT)[48] using discordant read pairs (DPs) to define potential MEI sites and split reads (SRs) to identify breakpoints and target site duplications (TSDs). MEIs initially were detected across all Phase 3 low coverage binary alignment/map (BAM) files generated in Phase 3 of the 1000 Genomes Project. Several samples with less than than 90% properly mapped read pairs were removed from the analysis because high levels of mapping artefacts in these samples confounded MEI detection. A total of 16,684 MEI insertions – 12,786 Alu, 3,060 L1, and 838 SVA (SINE/VNTR/Alu composite element) insertions – were identified from the remaining 2,453 low coverage samples.

4.3.9 Dinumt

SV Type(s): NUMT
Authors: Gargi Dayama, Ryan Mills

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/dinumt/`

Numt call-set generation: Nuclear insertions of mitochondrial DNA (numts) were discovered using dinumt[49] (version 0.0.22) in 1000 Genomes Project Phase 3 samples using the following parameters: `--len_cluster_include` = mean + 3 × standard deviation of sample insert size, `--len_cluster_link` = 2 × len_cluster_include, `--max_read_cov` = 5 × mean sample coverage. When possible, soft clipped reads were used to identify breakpoint positions. Confidence intervals were set to the distance between most prevalent clipped positions, if available; otherwise, confidence intervals were set to the inner distance between supporting read pair clusters. Calls

were then filtered based on the following criteria: phred-scaled quality filter <50, number of supporting reads <4, manual inspection, and then genotyped across the entire sample set.

5 Creation of the integrated callset

In this section, we detail the steps leading to the creation of a high-quality haplotype callset by integrating the sequencing data and the high-density microarray data available on the same samples. In brief, we began by creating a highly accurate haplotype scaffold from the microarray data by leveraging family information. From the sequence data, we calculated per-site, per-sample genotype likelihoods at a union set of the variants called in section 4. These likelihoods were then used to impute and phase onto the haplotype scaffold to create and final integrated callset.

For short variants, genotype likelihoods were calculated using one of two methods:

1. As described Section 5.1, SNPtools was used to generate the genotype likelihoods for pure biallelic SNPs
2. As described Section 5.2, glia and freebayes were used to generate genotype likelihoods for all variation which was putatively non-biallelic, non-SNP.

These results were subdivided as follows for input to either the joint imputation and phasing with Beagle/SHAPEIT2 (Section 5.4.2) or MVNcall (Section 5.4.3).

1. A high quality subset of the pure biallelic SNPs from the SNPtools pipeline (1) above which did not appear to overlap any other kind of SNP or indel variation in any of the input call sets and that passed an Support Vector Machine (SVM) filter controlling for FDR. This set of variants was included in the input to Beagle/SHAPEIT2.
2. Biallelic indels from the glia/freebayes pipeline (2) above with frequency of at least 0.5% that passed an SVM filter. This set of variants was included in the input to Beagle/SHAPEIT2.
3. Other variants, including low-frequency or low-quality indels, multiallelic SNPs, complex variants (mixtures of SNPs and indels in the form of block substitutions), and SNPs which had appeared as multiallelic but typed as biallelic by this process, were then used for integration with MVNCall (see Section 5.4.3).

For STRs, only the lobSTR callset (Section 4.2.1) was included in imputation and phasing with MVNcall. However, all STRs were removed from the final integrated release due to concerns about the FDR of this set (see Section 4.2.1).

For SVs, as described in 5.3, a merged set of high quality deletion calls were included in the input to Beagle/SHAPEIT2, while the remaining variants were sent to MVNcall.

Further details about these steps can be found in the sections below.

5.1 Generation of biallelic SNP genotype likelihoods

5.1.1 Generation of the union SNP allele list

Authors: Hyun Min Kang, Adrian Tan, Erik Garrison

Two separate processes were used to generate the list so as to minimize the likelihood of propagating method-specific issues into the union list and maximize the sensitivity of the resulting set. One, implemented at Boston College, generated a union list of SNPs by breaking apart any complex alleles (haplotype calls) from each caller and then removing duplicate SNPs and merging multiallelic SNPs into one record

`https://github.com/ekg/1000G-integration/blob/master/scripts/union/20130723_phase3_wg/union.snps.construction.zsh`

The other, implemented at the University of Michigan, focusing on SNPs only, decomposed multiallelic SNPs into biallelic SNPs, and created a union SNP list based on the base position and the non-reference allele. This SNP list was used for filtering biallelic SNPs.

5.1.2 Generation of biallelic SNP genotype likelihoods

Authors: Zhuoyi Huang, Fuli Yu

We calculated genotype likelihood (GL) of biallelic SNP using SNPTools[20]. We took the Phase 3 union list of putative SNP sites (section 5.1.1) from 2,535 samples and 80 TB low coverage and exome BAM files as input of GL calculation. The union list contains 95,472,850 SNPs with an average Ts/Tv ratio of 1.87.

In low coverage sequencing, insufficient evidence of alternate alleles and variability

of mapping and base quality may affect the accuracy of GL estimation. SNPTools employs the BAM-specific Binomial Mixture Model (BBMM) algorithm that overcomes the data heterogeneity due to sequencing platforms difference, reference bias and low data quality. For each sample, it fits the parameters of binomial models for different genotypes using all putative SNPs sites in both low coverage and exome BAM files of that sample, and calculates GL at each putative SNP site using the best-fit models. Finally the GL of each sample at all putative sites are aggregated into a population level GL VCF file.

We assessed the quality of Phase 3 genotype likelihood using the genotype information in the OMNI SNP array data as gold standard, which contains 2,141 samples. In a subset of 1,668 samples shared between 1000G Phase 3 and OMNI, we derived the genotype from the maximum genotype likelihood, and compared the genotype with that in the OMNI SNP array data. We obtained high genotype concordances, with 98.77% for REF/REF, 90.01% for REF/ALT and 99.00% for ALT/ALT.

We used the same cloud deployment of SNPTools pipeline (section 4.1.1), and performed the genotype likelihood calculation of 2,535 samples in the cloud using Amazon Elastic Compute Cloud (EC2) spot instances and the Elastic MapReduce (EMR) algorithm for distributed computation. The cloud processing of Phase 3 biallelic SNP GL calculation took 5 days.

5.1.3 Filtering of biallelic SNPs

Author: Hyun Min Kang

From the union SNP allele list generated at the University of Michigan, the SVM filtering implemented in the GotCloud was applied. In addition to the default variant features in GotCloud's filtering pipeline, the number of call sets that include the variants were added as an additional feature, and SVM filtering was applied.

The initial filtering results remove many rare variants (20% of singletons and 16% of doubletons), so we adjusted the SVM scores after evaluating false positives and false negatives in the following way. First, we performed the same SVM variant filtering procedure as above, but only focusing on the exome-target region, using only low-coverage genome sequences. Second, we genotyped the union sites within the exome-target region using the exome sequencing data only. Third, we annotated the union sites as true positive or false positives in the following way. We considered the variant as true positive if at least one exome-sequenced individual has depth of 30 or greater and has non-reference allele in the same sample where low-coverage genome reported non-reference allele. We considered the variant as false positive

if the variant is monomorphic in the exome sequence call set and has depth of 30 or greater in any of the samples where low-coverage genome reported non-reference allele. Next, we based on this annotations, we evaluated the false positive and false negative rate, stratified by non-reference allele count. Finally, based on this experiment, we adjusted the SVM score to have 5% FDR for rare variants. The adjusted SVM score was [(Original SVM Score) + 0.2 × (8 - min(Non-Reference Allele Count,8))]. As a result, the estimated FDR for singleton was 2.6%, and the fraction of removed singletons were reduced from 20% to 6.3%.

All pass variants were forwarded to the Beagle/SHAPEIT2 step (see Section 5.4.2) for phasing and imputation onto the haplotype scaffold.

5.2 Generation of non-biallelic, indel and complex genotype likelihoods

5.2.1 Generation of the union complex allele list

Author: Erik Garrison

The output of all SNP and indel-generating callers (section 4.1) was merged into a union allele list for both SNPs and indels (`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/union/ALL.wgs.union_from_bc.20130502.snps_indels_complex.sites.vcf.gz`). This set was created by first running `vcfallelicprimitives` (`https://github.com/ekg/vcflib/blob/master/src/vcfallelicprimitives.cpp`) on each callset, merging them using a series of calls to vcfintersect -u (`https://github.com/ekg/1000G-integration/blob/master/scripts/union/20130723_phase3_wg/genunion.zsh`) and then applying the following normalisation process (described `https://github.com/ekg/1000G-integration/blob/master/scripts/union/20130723_phase3_wg/cleanunion.zsh`):

- left alignment via `vcfleftalign`
- simplification and complex allele decomposition using `vcfallelicprimitives`
- normalization into single-record-per-line format removal of duplicate alleles (`vcfstreamsort`, `vcfuniq`)
- generation of multiallelic records for overlapping variants (`vcfcreatemulti`)

5.2.2 Generation of non-biallelic, indel and complex genotype likelihoods

Authors: Erik Garrison, Shane A. McCarthy, Xiangqun Zheng-Bradley

Genotype likelihoods were generated across this set of candidate alleles by a novel process (glia) wherein the project alignments were realigned to a graph constructed from the candidate alleles in the union allele set. Were the alignment to the graph had fewer mismatches, gaps, or soft clips than the project alignment, the new alignment replaced the old for the generation of genotype likelihoods. Genotype likelihoods were calculated using freebayes v0.9.9.2-26-g8a98f11, with default parameters except the use of `--min-repeat-entropy 1`, which ensures that calls were made over sequence with entropy greater than 1.0, per-read-group `--contamination-estimates` (`https://github.com/ekg/1000G-integration/blob/master/resources/p3.exome_lowcov.per_RG.het_and_contam.contaminations`) provided by VerifyBamID, and the union allele list as the `--haplotype-basis-alleles`. Males were called as haploid on the non-PAR regions of chrX. The genotype likelihood generation process is described in a script that was used by the centres that participated in the generation of these likelihoods (`https://github.com/ekg/1000G-integration/blob/master/scripts/run_region.sh`). To complete the work in a timely manner, this genotype likelihood generation method was run at the Sanger, EBI, and Boston College.

5.2.3 Filtering of indels

Authors: Hyun Min Kang, Erik Garrison

Biallelic indels with frequency of at least 0.5% that passed an SVM filter developed by the University of Michigan were merged with the SNPtools biallelic SNPs for integration with Beagle and SHAPEIT2 (see Section 5.4.2). The University of Michigan indel filters were obtained from the SVM filtering implemented in GotCloud, by adding two features – the number of callsets that includes the variant, and the frequency of flanking 2-mer, 3-mer, and 4-mers compared to their distribution across all other variants – were added in addition to the default feature prior to perform SVM calling.

Other variants, including low-frequency or low-quality indels, multiallelic SNPs, complex variants (mixtures of SNPs and indels in the form of block substitutions), and SNPs which had appeared as multiallelic but typed as biallelic by this process, were then used for integration with MVNCall (see Section 5.4.3).

5.3 Merging SV callsets

In order to generate a high confidence set of large deletion sites to be used for joint haplotype scaffold generation along with SNPs and indels, we employed GenomeSTRiP[46] to re-genotyping sites called with the five most specific deletion discovery algorithms (BreakDancer[40], DELLY[41], CNVnator[43], GenomeSTRiP[46], and VariationHunter[42]). GenomeSTRiP's redundancy removal function was used to merge these sites into a coherent list of large high confidence deletions. To further reduce redundantly called sites, a second round of redundancy removal was performed using a more stringent protocol.

This list was used for haplotype scaffold generation, along with SNPs and biallelic indels, using SHAPEIT2 (see Section 5.4.2).

All other SV callsets were phased into these haplotype scaffolds using MVNcall (see Section 5.4.3).

5.4 Genotype calling and estimation of an integrated set of haplotypes

Authors: Olivier Delaneau, Androniki Menelaou, Shane A. McCarthy, Hyun Min Kang, Erik Garrison, Jonathan Marchini

Genotypes and haplotypes at all SNPs, indels and structural variants were called using a 3 step process that aimed to integrate the project's sequencing data with microarray genotypes available on the same samples. Microarray genotypes on the project samples were used to create a highly accurate haplotype scaffold by leveraging family information. All remaining sequenced sites were then phased onto this scaffold. This overall strategy has been shown to produce low error rates for genotype calls, and a set of haplotypes that produce good performance when used as a reference panel for downstream imputation into GWAS cohorts[50].

5.4.1 Creation of a haplotype scaffold from microarray genotypes

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/`

All 2,535 sequenced samples were genotyped on either the Illumina Omni 2.5 or Affymetrix 6.0 microarray, as well as an additional set of 2,322 unsequenced sam-

ples, many of whom are close relatives of the sequenced samples. The Illumina Omni 2.5 contains 1,722 sequenced and 596 unsequenced samples that form 49 duos, 403 trios and 1,011 unrelateds. The Affymetrix 6.0 contains the remaining 813 sequenced samples plus another set of 1,673 samples and forming 102 duos, 416 trios and 1,034 unrelateds (Supplementary Information Table 11). SHAPEIT2[51] was used to estimate haplotypes from the 2,318 IlluminaOmni2.5 samples, and the 2,486 Affymetrix6.0 samples, in separate runs of the program. SNPs with a missing data rate above 10% and a Mendel error rate above 5% were removed before phasing. All other Mendel errors were set to missing and genotypes were imputed as part of the phasing. After this filtering the OMNI2.5 and Affymetrix6.0 datasets consisted of 2,083,066 and 873,696 SNPs respectively. SHAPEIT2 was run using the following settings (W = 2Mb, K = 200 haplotypes, burnin iterations = 10, pruning iterations = 10, sampling iterations = 50). The pseudo-autosomal regions of the X chromosome were processed separately to the non-pseudo-autosomal region.

5.4.2 Joint phasing of biallelic SNPs, high-confidence indels and large deletions onto the haplotype scaffold

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/genotype_likelihoods/shapeit2/`

We took genotype likelihoods at 80,106,118 biallelic SNPs (Section 5.1), 986,877 high-confidence biallelic indels (Section 5.2) and 34,171 deletions (Section 5.3) and phased these sites onto the haplotype scaffolds produced above. This set of variants was deemed to be a high-quality set of sites that could be phased jointly.

Firstly, BEAGLE4[52] was run to obtain an initial set of genotypes and haplotypes. Beagle was run on chunks of 12,000 sites, with an overlap of 2000 sites between consecutive chunks. Beagle was run using 5 burn-in and 15 sampling iterations.

SHAPEIT2 was used to phase the genotype likelihoods onto the haplotype scaffold. Genotypes called by Beagle with a posterior probability greater than 0.995 were fixed as known genotypes. In addition, the haplotypes estimated by Beagle were used to initialize the SHAPEIT2 phasing. SHAPEIT2 was run with 12 pruning stages of 4 iterations. These iterations gradually reduce the complexity of the per-individual Markov chains used to model the space of haplotypes consistent with each sample. This was followed by 20 sampling iterations, that were used to estimate the final set of haplotypes. The window parameter in SHAPEIT2 was set at 0.1 Mb (–window 0.1), the number of Hamming distance conditioning haplotypes was set to 400 (–states 400), and the number of random conditioning haplotypes was set to 200 (–states-random 200). SHAPEIT2 was run in chunks of 1.4 Mb with 0.4 Mb overlap between

successive chunks. Since all sites have been phased onto a chromosome-wide haplotype scaffold, the SHAPEIT2 haplotypes from each chunk can be ligated together simply by concatenating the chunks (minus the overlapping regions) together.

5.4.3 Phasing of all other sites onto the scaffold

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/genotype_likelihoods/mvncall_input/`

For all remaining variants, we used the program MVNcall[53] to phase the genotype likelihoods of one site at a time onto a haplotype scaffold. This set of sites consisted of a total of 12,208,480 indels, multiallelic SNPs, complex and structural variants, and Short Tandem Repeats (STRs). This step has the advantage that sites can be phased without affecting the phasing of the high-confidence set of sites from Step 2, and is highly parallelizable.

The haplotype scaffold used for this stage was derived from the haplotypes produced in Step 2, by extracting only 1,992,184 sites present on the OMNI2.5 chip, while maintaining the haplotype order. For the phasing of the variants using MVNcall, the following parameters were used (conditional haplotypes –k=100, iterations –iterations=40 and burn-in iterations –burnin=20). The remaining parameters were set to their default values. All samples were processed together.

The initial publication of MVNcall[53] discussed only the phasing of biallelic variants. The method is now extended to handle multiallelic variants. Biallelic sites are commonly coded as '0' for the reference allele (r) and '1' for the alternative allele (a). We extended this representation to code multiallelic sites, as a form of multiple biallelic sites. A multiallelic site with k alternative alleles was coded as $k-1$ biallelic variants. In this way, haplotypes across these $k-1$ sites code for alleles. The ith alternate allele is coded as a haplotype with a 1 at the ith biallelic site and 0's at the $(k-2)$ other biallelic sites. In the MCMC algorithm the resulting Gibbs sampling updates involve multi-variate Normal densities (of dimension $k-1$). When jointly sampling new alleles across these $k-1$ biallelic sites, only the ${}^{k-1}C_2$ combinations of haplotypes are considered.

In addition, MVNcall was extended to handle phasing on the non-pseudo-autosomal region of chromosome X (option –chrX). It assumes the input VCF on the non-pseudo-autosomal region includes only the homozygous genotype likelihoods for the males and all the possible genotype likelihoods for the females. The pseudo-autosomal regions of chromosome X were processed separately to the non-pseudo-autosomal region. The parameter settings used for the autosomal chromosomes were

also used for chromosome X.

5.5 Final filtering

5.5.1 Filtering of non-biallelic, non-SNP variants

Author: Erik Garrison

The MVNCall results were filtered in a two-pass approach. In the first pass, genotyping was carried out over the full set of alleles. This produced genotype posterior quality estimates which we observed to be strongly correlated with genotype accuracy and also allele call accuracy. These were averaged at each site and combined with other allele-level annotations derived from information provided by freebayes in the genotype likelihoods from the second process described above. This set of information was merged with confident genotype calls on 24 samples which we had sequenced to high depth using long (2×250 bp) Illumina reads, and the genotype concordances between the post-imputation low-coverage data and the high-coverage samples were used to train an SVM filter based on the annotations. A set of filters were determined so as to maintain <5% FDR for all allele classes (see section 6.1), and this was used to filter the set on a per-allele basis. The results were then imputed again using MVNCall, yielding genotypes for each sample across the confident alleles in the set.

5.5.2 Filtering of structural variants

Author: Jan Korbel

We performed another SV merging and filtering step to remove redundant calls, to harmonize the SV notation and to ensure a site FDR <5% for the merged SV call set. All post-phasing mono-monomorphic reference sites were excluded, cryptic related samples were dropped and CNVs were classified as biallelic deletions (DEL), biallelic duplications (DUP) and multiallelic copy-number variants (mCNV). Merging was performed using an overlap graph $G(r, c) = G(0.71, 0.71)$, requiring a reciprocal overlap (r) of at least 71% and a non-reference copy-number concordance (c) of at least 71%. Using these cutoffs ensured that >99% of all connected components in the overlap graph were cliques. For each connected component, we picked one representative call whereas all merged calls were specified in the VCF INFO column.

5.6 Integration of phasing results and generation of final released haplotype set

Authors: Erik Garrison, Xiangqun Zheng-Bradley, Laura Clarke

The resulting SHAPEIT2 and MVNCall genotype sets were merged using the following approach. The allelic representation of both sets was normalized using vcfallelicprimitives and then vt normalize. Genotypes for indels and SNPs were taken from the normalized SHAPEIT2 results except where they overlapped a multiallelic locus from the normalized MVNCall set. At these sites, the MVNCall genotypes were passed forward and the SHAPEIT2 ones were suppressed. The final merge step is described in the following script: (`https://github.com/ekg/1000G-integration/blob/master/scripts/merge/get_final_merged_region.zsh`)

From the 2,535 samples processed to this point, we removed the genotypes of 31 individuals who have a blood relationship with remaining 2,504 samples in the main release. This was done to ensure we do not over estimate allele frequency. The main release VCF[54] files are available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`

The 31 related samples are listed in

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/20140625_related_individuals.txt`

This has resulted in a small number of AC=0 sites from rare alleles only present in one or more of these 31 individuals. The genotypes for 31 individuals are available here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related_samples_vcf/`

5.7 Phase 1 variants not in Phase 3

Authors: Xiangqun Zheng-Bradley and Laura Clarke

The Phase 1 variant list released in 2012 and the Phase 3 variant list released in 2014 overlap but Phase 3 is not a complete superset of Phase 1. The variant positions between the Phase 3 and Phase 1 releases were compared using their positions. This shows that 2.3M Phase 1 sites are not present in Phase 3. Of the 2.3M sites, 1.92M

are SNPs, the rest are either indels or structural variations (SVs).

The difference between the two lists can be explained by a number of different reasons.

1. Some Phase 1 samples were not used in Phase 3 for various reasons. If a sample was not part of Phase 3, variants private to this sample are not be part of the Phase 3 set.

2. Our input sequence data is different. In Phase 1 we had a mixture of both read lengths 36 bp to >100 bp and a mixture of sequencing platforms, Illumina, ABI SOLiD and LS454. In Phase 3 we only used data from the Illumina sequencing platform and we only used read lengths of 70 bp+.

Reasons 1 and 2, the absent samples and non-Illumina data can explain 548K missing SNPs, leaving 1.37M SNPs still to be explained.

3. The Phase 1 and Phase 3 variant calling pipelines are different. Phase 3 had an expanded set of variant callers, used haplotype aware variant callers and variant callers that used de novo assembly. It considered low coverage and exome sequence together rather than independently. Our genotype calling was also different using SHAPEIT2 and MVNcall, allowing integration of multi allelic variants and complex events that weren't possible in Phase 1.

   891K of the 1.37M sites missing from Phase 1 were not identified by any Phase 3 variant caller. These 891K SNPs have relatively high Ts/Tv ratio (1.84), which means these were likely missed in Phase 3 because they are very rare, not because they are wrong; the increase in sample number in Phase 3 made it harder to detect very rare events especially if the extra 1,400 samples in Phase 3 did not carry the alternative allele.

   481K of these SNPs were initially called in Phase 3. 340K of them failed our initial SVM filter so were not included in our final merged variant set. 57K overlapped with larger variant events so were not accurately called. 84K sites did not make it into our final set of genotypes due to losses in our pipeline. Some of these sites will be false positives but we have no strong evidence as to which of these sites are wrong and which were lost for other reasons.

4. The reference genomes used for our alignments are different. Phase 1 alignments were aligned to the standard GRCh37 primary reference including unplaced contigs. In Phase 3 we added EBV and a decoy set to the reference to reduce mismapping. This will have reduced our false positive variant calling as it will have reduced mismapping leading to false SNP calls. We cannot quantify this effect.

We have made no attempt to eludcidate why our SV and indel numbers changed. Since the release of Phase 1 data, the algorithms to detect and validate indels and SVs have improved dramatically. By and large, we assume the indels and SVs in Phase 1 that are missing from Phase 3 are false positive in Phase 1.

More details about the comparison can be found here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/phase1_sites_missing_in_phase3/`

6 Validation

6.1 Validation and filtering of short variants

Authors: Anthony Marcketta and Adam Auton

As described above, two programs were used to genotype and phase variants in the low coverage sequencing data from 2,535 individuals: SHAPEIT2 and MVNcall. SHAPEIT2 was used to phase high-quality biallelic variants, which we expected to have high validation rates, while MVNcall was used for more complicated events. In order to ensure that the variants included within the callset met the required 5% FDR threshold, we adopted the following procedure.

First, variants were divided into five distinct categories: SHAPEIT2 biallelic variants, MVNcall SNPs, MVNcall MNPs, MVNcall indels, and MVNcall complex sites. For each of these categories, 20,000 autosomal sites that were called as variant within one of the 26 high-coverage non-PCR individuals were randomly selected for validation.

To validate alleles, we aimed to identify supporting reads within the high-coverage PCR free data. This was achieved by locally realigning the high coverage data around each putative allele using a Smith-Waterman local alignment algorithm with a high gap penalty parameter. For each randomly selected site, local alignment of reads was performed for all alternate alleles that were called within the 26 individuals by the main project (i.e. low coverage and exome) data. Only reads that overlapped the site by at least 35 bases on either side were considered.

In order to call a read as supporting a given allele, the a perfect match alignment was required within the allele region and at the flanking 1 base pairs. In addition, no more than one mismatch was allowed in the 10 flanking bases on either side, excluding sites known to polymorphic from Phase 1. No gaps were allowed within

the alignment within any of these regions. If a read met these criteria, it is considered as a read that supports the allele. Reads were not counted that matched multiple alleles. An allele was considered as validated it was supported by at least two reads within at least one PCR-free individual.

Having validated each allele, we next aimed to estimate the FDR for the combined dataset. As we were validating only within 26 individuals, a naive calculation would likely underestimate the FDR as it would be biased towards common variants. In order to provide a more accurate estimate, we calculated a weighted FDR that accounted for allele frequency. Specifically, if f_i represents the false positive rate for alleles (whole-sample) allele count i, then the weighted FDR for the whole dataset was estimated as $\sum_i w_i f_i$, where w_i represents the proportion of alleles in the complete dataset with allele count i.

Using this method, we established that the SHAPEIT2 variants met the 5% FDR requirement without further filtering. This was not the case for more complex types of variation that had been genotyped by MVNcall. In order to provide a filtering metric for the MVNcall alleles, we used a support vector machine (SVM), as described elsewhere. We then filtered the MVNcall alleles in order to achieve a FDR below while maintaining the largest possible number of alleles within the callset. This was achieved by only keeping alleles that met the following thresholds:

- Indels – an allele count of $\geq$3 and an SVM score of $\geq$0.67
- SNPs – an allele count of $\geq$2 and an SVM score of $\geq$0.78
- MNPs – an allele count of $\geq$2 and an SVM score of $\geq$0.64
- Complex variants – an allele count of $\geq$3 and an SVM score of $\geq$0.88

A similar validation method was used for candidate variants in the non-pseudo-autosomal region of chromosome X, with parameters altered to account for the reduced number of chromosomes included in the sample. For this analysis, 1000 candidate sites of each type were randomly selected for validation. The filtering thresholds needed to meet the 5% FDR requirement for alleles on chromosome X are the following:

- Indels – an SVM score of $\geq$0.59
- SNPs – none
- MNPs – an SVM score of $\geq$0.86
- Complex variants – an allele count of $\geq$2 and an SVM score of $\geq$0.51

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/high_coverage_alignments/`

6.2 Structural variant validation

We used different methodologies to assess the FDR of structural variants, including microarrays, PCR-free whole genome sequencing and PacBio sequencing. We estimate FDR<5% for deletions, duplications, multi-allelic copy-number variants, Alu insertions and L1 insertions, and FDR<20% for inversions, SVA insertions and NUMTs (see the SV companion paper for more details[39]).

Experimental conditions of PCR validations: PCR experiments were carried out in different laboratories, focusing on different SV types: European Molecular Biology Laboratory (DEL, DUP, INV, NUMT; section 6.2.1), Louisiana State University (MEI; section 6.2.2), and University of Michigan (NUMT; section 6.2.3).

Experimental conditions of PacBio validations: Long-read (PacBio) SMRT sequencing focussed on validation on inversions at University of Washington (INV; section 6.2.4).

6.2.1 European Molecular Biology Laboratory (DEL, DUP, INV, NUMT)

Authors: Adrian Stütz, Benjamin Raeder, Jan Korbel

PCR primers were obtained from Sigma, after primer design using an in-house pipeline based on BLAST[55] as well as primer3 software (`http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi`). PCR was preformed using 10 ng of genomic DNA (Coriell) in 25 μl volumes using the Sequalprep Long PCR reagents (Life technologies) in a 96 well plate using the DNA Engine Tetrade 2 thermocycler (BioRad). PCR conditions were: 94 °C for 3 min, followed by 10 cycles of 94 °C for 10 s, 62 °C for 30 s and 68 °C for 6 min and 25 cycles of 94 °C for 10 s, 60 °C for 30 s and 68 °C for 8 min, followed by a final cycle of 72 °C for 10 min. PCR products were analysed on a 0.8% agarose gel stained with Sybr Safe Dye (Life Technologies) and a 100 bp ladder and 1 kb ladder (NEB). If necessary, gel bands were cut with a scalpel, gel extracted with the Nucleospin Gel and PCR Cleanup kit (Macherey-Nagel) and send for capillary sequencing (GATC Biotech AG).

6.2.2 Louisiana State University (MEI)

Authors: Miriam Konkel, Jerilyn Walker, Mark Batzer

PCR reactions were performed under the following conditions using a standard Taq

polymerase: initial denaturation at 94 °C for 90 sec, followed by 32 cycles of denaturation at 94 °C for 30 sec, annealing at 57 °C and extension at 72 °C for 30 to 90 sec depending on the predicted PCR amplicon size. PCRs were terminated with a final extension at 72 °C for 2 min. For the amplification of the entire L1 using LA-Taq DNA polymerase, the above-described protocol was modified in the following way. The extension step of each cycle was carried out at 68 °C for 8 min 30 sec, followed by a final extension step at 68 °C for 10 minutes at the end of the run. All PCR products (20 μl) were size-fractionated in a horizontal gel chamber on a 2% or 1% (for loci amplified with LA-taq) agarose gel containing 0.1 μg/ml ethidium bromide for 45-60 minutes at 175-200 V or 1 hour/45 min at 150 V, respectively. DNA fragments were visualised with UV-fluorescence and images were saved using a BioRad ChemiDoc XRS imaging system (Hercules, CA).

6.2.3 University of Michigan (NUMT)

Authors: Sarah Emery, Jeffrey Kidd

Numts identified by computational analysis were validated by polymerase chain reaction (PCR) and Sanger sequencing of amplicon(s) that spanned 50-500 bp of gDNA flanking the insert, the breakpoint between the gDNA and the insert, and the insert. Primer sets that hybridize to the gDNA flanking the insert were designed using Primer3 Software (`http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi`) and amplification was done with Platinum Taq (Invitrogen Life Technologies, Gaithersburg, MD), Picomaxx (Agilent Technologies, Palo Alto, CA), or LongAmp (New England Biolabs, Beverly, MA) products in a 20-50 μl reaction volume containing 50 ng of template DNA, 1 uM primer, and 1.5 mM MgCl2 if not supplied in the PCR buffer. Thermocycling was done for 30 cycles at 56-67 °C annealing temperature and 1-15 minute extension time. For inserts less than 3 kb, a PCR product of the predicted size was identified in individuals homozygous or heterozygous for the insert by agarose gel electrophoresis and the insert was sequenced in one individual. Amplicons of interest were purified from a PCR reaction for homozygous individuals (Qiaquick PCR purification kit, Qiagen, Valencia, CA) or isolated from the gel for heterozygous individuals (Qiaquick Gel Extraction Kit, Qiagen) and sequenced at the University of Michigan Sequencing Core. For inserts larger than 3 kb, a PCR product of the predicted size was identified in individuals heterozygous for the insert by gel electrophoresis. For sequencing, two overlapping PCR products were made using primer sets designed as outlined above with one primer that binds in the gDNA flanking the insert and one primer that binds in the middle of the insert.

6.2.4 University of Washington (INV)

Authors: Maika Malig, Mark Chaisson, Evan Eichler

We selected a total of 35 inversion sites (inferred by DELLY) from two genomes (NA12756 and NA19129) for validation using long-read (PacBio) SMRT sequencing of fosmid clone inserts (∼40 kbp). A total of 113 clones (2-4 clones per site) were selected and grown based on mapping of fosmid end-sequence pairs to GRCh37[56]. DNA was individually prepared for each clone (High Pure Plasmid Isolation KitTM, Roche) and DNA from 7-8 clones were pooled. A 20 kbp SMRTbellTM template library was prepared for each pool; the library sequenced with one SMRTcell per pool using either P4-C2 or P5-C3 chemistry and inserts were assembled using HGAP and QUIVER post-processing[57] as previously described[58]. 111/113 (98.2%) of the clone inserts resolved into a single sequence contig with on average 400-fold sequence coverage per fosmid clone insert. Assemblies were compared with GRCh37 using Miropeats[59] and dotplot analysis to identify breakpoints and confirm inversion status. Overall, 82.3% (28/34) of sites validated with 1 site excluded due to sequence complexity. This is a conservative estimate because only one haplotype was recovered for 2/6 of the invalidated sites. Excluding these two sites would result in a validation rate of 87.5%. We further employed PacBio reads from the recently sequenced CHM1 genome[58] for the verification of Phase 3 inversions which the EMBL group genotyped into CHM1 using published CHM1 Illumina sequencing data[60].

See the companion paper[39] for more details.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/inversion_validation/20150511_inversion_validation_S15.xlsx`

6.3 SNP haplotype validation by comparison with phased haplotypes obtained from fosmid pool sequencing

Authors: Shiya Song, Elzbieta Sliwerska, Sarah Emery, Jeffrey M. Kidd

6.3.1 Data Production

Genomic DNA for fosmid library construction from samples NA19240, HG03428, HG02799, and HG03108 was purchased (Coriell, Camden, NJ) or isolated from cell lines (Coriell) using Gentra Puregene Blood Kit (Qiagen, Valencia, CA). Aliquots of 10 μg of DNA were sheared in 120 μl volumes on a Digilab Hydroshear for 60 cycles

at a speed code of 16 or 20 cycles at a speed code of 20. Sheared DNA was loaded and ran on a pulse field gel at 200 V for 26 hours with 0.5-15 s switching or a BioRad CHEF DR III (Hercules, CA) at 6 V/cm for 16 hours with 1-6 s switching. DNA from 25-45 kb was cut out of the gel and isolated by electroelution for 12 hours at 120 V or 3 hours at 150 V. After electroelution, DNA was isolated with Ampure XP beads, end-repaired with the Epicentre End-It kit and ligated to the Epicentre pCC1Fos fosmid arms. The resulting ligation was packaged and transfected into the Phage T-1 Resistant EPI300-T1 E. coli plating strain (Catalog Number CCFOS110). One hour after transfection, the resulting cells were split into the appropriate volumes to give pools of 1,500-3,000 cells per pool. Barcoded libraries for sequencing were constructed from mini-prepped DNA obtained from each pool using either the Illumina Nextera or Bioo NEXTFlex protocols. In addition, we created high-coverage (>20×) coverage of standard Illumina whole genome sequencing. In addition, we re-processed fosmid clone pool data from GIH sample NA20847[61] (SRA026360) and directly used phased haplotypes for CEU sample NA12878 obtained from Duitama *et al.*[62] (ERP000819).

Sample	Population	Number of Clones After Filter	SRA Accession
NA19240	YRI	521,783	SRS628777
HG03428	MSL	1,424,234	SRS722908
HG02799	GWD	1,141,020	SRS722940
HG03108	ESN	1,058,027	SRS722941

6.3.2 Fosmid clone identification and haplotype construction

We implemented a method to detect fosmid clone boundaries based on pooled sequence data (Song *et al.*, in preparation). Briefly, reads from each clone pool were mapped to a reference assembly including the human genome (GRCh37/hg19), Epstein Barr virus, the E. coli genome and the fosmid vector backbone using BWA v0.5.9-r16. Candidate fosmid clones were identified by computing read-depth in 1 kbp windows for each clone pool and merging consecutive windows allowing a maximum gap of 3 windows. Reads where one end mapped to the fosmid vector backbone and another end mapped to human genome, called anchoring reads, were used to better assign clone breakpoints. Observing anchoring reads in the middle of consecutive windows identified overlapping clones, which were excluded from downstream analysis. Each clone pool was separately genotyped at heterozygous SNPs called from high-depth whole genome shotgun sequencing using GATK UnifiedGeno-

typer v2.3-9. Clones covering one or more heterozygous SNP positions detected in the WGS data were used to resolve haplotypes in next stage. A small proportion of clones (8.1% for NA19240 as example) were genotyped as heterozygous, probably resulting from overlapping clones or mapping errors. These clones were excluded from further analysis. We applied RefHap[62] to the resulting data to create phased haplotype blocks. For samples NA19240, HG02799, and HG03108, we further compared with phased SNP genotypes obtained from trio genotyping. Using these data, we corrected switch errors within RefHap computed blocks, and assigned parental alleles within each block, thus linking adjacent blocks together to produce near-to-complete haplotypes. For HG03428 and NA20847, trio data was unavailable and phasing comparison analysis was limited to comparisons within RefHap blocks.

Sample	Pop.	Number of clones after filter	Number of RefHap Blocks	N50 of phased blocks (kbp)	SNPs to be phased	% phased SNPs
NA19240	YRI	521,783	16,334	347	2,588,454	92.70%
HG02799	GWD	1,141,020	5,236	1416	2,780,269	98.40%
HG03108	ESN	1,058,027	5,416	1294	2,756,725	98.40%
HG03428	MSL	1,424,234	4,390	1849	2,775,099	99.30%
NA20847	GIH	571,419	16,838	385	1,680,704	93.37%
NA12878	CEU	–	–	–	1,843,256	82.90%

6.3.3 Haplotype comparison

We compared 1000 Genomes Phase 3 haplotypes with haplotypes obtained using fosmid pool sequencing. Switch error is an inconsistency between an assembled haplotype and the real haplotype between two contiguous variants. Switch error rate is switch error normalized by number of variants for comparison. Overall, the haplotype concordance between 1000 Genomes Phase 3 haplotypes and physically phased haplotypes are quite high, 96.42% in average, with switch error rate around 0.56% and mean inter-switch distance 1062.1 kbp. Among switch errors, 85.7% are flip errors, namely individual alleles appearing on the opposite haplotype, indicating overall high quality of long range haplotypes. These results are shown in Supplementary Information Table 12.

7 Chromosome Y integrated callset

Out of the 2,535 individuals in Phase 3, there were 1,244 males. The variant calling and integration process for chromosome Y on these male samples followed a related, but different path from that of the autosomes and chromosome X. Only a subset of the variant callers made callsets available for chromosome Y. Genotype calling from the union site list followed that used for indels in the main callset (section 5.2), while the inferred phylogeny was used to impute missing genotypes and assign ancestral alleles. Most of these details are left to the companion manuscript[63], however we give a brief overview here.

7.1 Short variant callset

Authors: G. David Poznik, Shane A. McCarthy, Juan L. Rodriguez-Flores, Yali Xue and Chris Tyler-Smith

SNP, indel and complex variant calling was confined to the 10.3 Mb of the Y chromosome within which one can reliably call genotypes using short-read sequencing[64]. These regions are defined in this BED file:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/chrY_callable_regions.20130802.bed`

Short variant callsets that were made for chromosome Y are listed here along with any notes about differences from details in Section 4.1.

- **Boston College – Freebayes**: See Section 4.1.2. Ploidy was set for males to 1 and females to 0 using the '–cnv-map' option.

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/bc/ALL.chrY.bc.20130502.snps_indels_mnps_complex.sites.vcf.gz`

- **Oxford – Platypus**: See Section 4.1.6.

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/ox/ALL.chrY.oxford_platypus.20130502.snps_indels_mnps_cplx.low_coverage.genotypes.vcf.gz`

- **Oxford – Cortex**: See Section 4.1.7.

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/20130502/supporting/input_callsets/cortex/ALL.wgs.cortex.20130502.biallelic_snps_indels.low_coverage.sites.vcf.gz`

- **Sanger Institute – SAMtools/BCFtools**: See Section 4.1.8. SAMtools likelihoods and site statistics were generated based on all male and female sample BAMs, but BCFtools variant calling was restricted to haploid calls on the 1,244 male samples with the '-S' and '-Y' options.

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/si/ALL.chromY_uniq.samtools.20130502.snps.integrated.genotypes.vcf.gz`

- **Sanger Institute – SGA-Dindel**: See Section 4.1.9.

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/si/ALL.chrY.sga.20130502.snps_indels_mnps.low_coverage.genotypes.vcf.gz`

Two separate GATK UnifiedGenotyper (see section 4.1.3) SNP callsets were made: one at Stanford and one at Cornell.

- **Stanford – GATK UnifiedGenotyper**

  GATK (2.5.2) UnifiedGenotyper was run jointly across all male samples to call SNPs within Y:2,600,000-28,800,000 using very liberal thresholds: –min_base_quality_score 17, -stand_call_conf 4.0, -stand_emit_conf 4.0. Conditional on GATK identifying the site as variable, genotypes were called according to maximum likelihood, setting the genotype to missing whenever the Phred-scale likelihood difference was less than 20 or there was a heterozygous maximum likelihood genotype. Next, site-level SNP filters were applied.

  - Filtered depth: 2,000–6,000 ($\sim$6 MAD interval)
  - MQ0 Ratio: $\leq$0.1
  - Missingness: $\leq$400 individuals ($\sim \frac{1}{3}$ of sample)
  - Heterozygous Maximum Likelihood: $\leq$200 individuals
  - Missingness OR Heterozygous Maximum Likelihood: $\leq$400 individuals

  `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/stn/ALL.chrY.stanford_v1.20130502.snps.low_coverage.genotypes.vcf.gz`

- **Cornell – GATK UnifiedGenotyper**:

  Genotype calls were generated simultaneously for all male samples using the GATK UnifiedGenotyper v2.4, allowing all variant sites Q30 or higher. Diploid calls for variant sites were subsequently re-evaluated, marking homozygous reference '0/0' genotypes as haploid reference '0', heterozygous '0/1' genotypes

as missing '.', and homozygous alternate '1/1' genotypes as haploid alternate '1'. Genotypes where depth was below 2x were also marked as missing '.'. Sites with two or more variant alleles were excluded, as were sites with variant allele count = 0 after filtering, and sites with a genotype call rate below 50%. After conversion of genotypes to haploid, INFO tags were re-calculated, including AC (variant allele count), AF (variant allele frequency), AN (allele number), SM (sample count), ANP (proportion of non-missing genotype calls), and DP (depth). In addition, an INFO tag marking variants within 10 Mb of ChrY ideal for next generation sequencing and population genetic analysis were marked with the 'POZNKIK13' INFO flag. A total of 150,379 bi-allelic SNPs were identified (Ts/Tv=1.32), including 83,087 within the defined 10 Mb intervals.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/cornell/ALL.ChrY.Cornell_v2.20130502.SNPs.Genotypes.vcf.gz`

For the short variant callsets above, a union site list was made and fed into the genotype likelihood process described in Section 5.2.2.

7.1.1 Filtered SNP callset and phylogentic tree

The set of biallelic SNPs was extracted from from the union genotype likelihoods. We then imposed six filters, restricting to (a) biallelic SNPs with (b) genotype quality (QUAL) greater than one; (c) filtered read-depth across all samples in the range 2000–6000 (1.6×–4.8×), which represents the median depth across sites, plus or minus three median absolute deviations; (d) no more than 10% of reads with mapping quality scores of zero; (e) no more than 400 samples (∼one third of the total) with no high quality reads mapping to the site; and (f) no more than 200 samples whose maximum likelihood genotype state was heterozygous.

Upon conducting a phylogenetic analysis and assigning SNPs to branches of the phylogeny, we observed that the genotype calls in this preliminary consensus call set were marred by reference bias. A greater than expected proportion of sites were incompatible with the phylogeny. These incompatibilities were often traceable to either reference genotype calls where read data were contradictory and a no-call would have been most appropriate or to cases where the read data supported a non-reference genotype call but were not sufficient to surmount the strong prior induced by over 1,000 reference genotype calls in the sample. Therefore, at each site, we replaced the FreeBayes calls by the maximum likelihood genotype state for each sample, subject to the condition that the likelihoods for reference and non-reference states differed by two log units. When the absolute difference in likelihoods was less than or equal to two log units, we assigned a no-call.

This approach yielded a genotype call set of 59,675 variable SNPs. We then identified additional biallelic SNPs by splitting complex sites into biallelic components using 'bcftools norm'. We applied identical filters and added the remaining 880 sites to the final call set, numbering 60,555 SNPs.

Haplogroup assignments are in this file:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/1000Y.sample_level_haplogroup_calls.ver4b.txt`

7.1.2 Imputing missing genotypes and identifying ancestral allele states

We leveraged the inferred phylogeny to impute missing genotypes and infer ancestral allele states[64]. Prior to doing so, we partitioned the tree into eight overlapping subtrees in order to partially account for homoplasy and to reduce the impact of genotyping error. For each subtree, we defined a set of sites that were variable and assigned each site to the internal branch constituting the minimum superset of carriers of one allele or the other. Let t represent this minimum superset. We designated the derived state to the allele that was observed only within t and the ancestral state to the other allele. When the dichotomy was clean (i.e., no ancestral alleles were observed within t), we deemed the site compatible with the subtree and imputed missing genotypes accordingly. Otherwise, for sites incompatible with the subtree, we did not impute missing genotypes.

This imputation procedure works well for well-balanced trees, where the superposition of lineages elicits high effective coverage on the internal branches of the tree. It breaks down, however, in instances where the outgroup of a clade is represented by just one or two low-coverage samples. When an outgroup lacks data for a given site, the site cannot be assigned to the branch immediately upstream of the outgroup. Instead, these sites will be misassigned to the branch one level downstream. Therefore, ancestral allele imputations must be interpreted with caution for such lineages.

By including a small set of samples that overlapped with neighbouring subtrees, we polarized ancestral and derived states for all branches but the most basal of the global phylogeny. We could not polarize SNPs mapping to the two branches separating hgA0 from the rest of the tree, as no outgroup was available for this most ancient split.

Due to reversion mutations, alleles that are ancestral in one subtree may be derived in another. We determined the globally ancestral allele based on the outermost subtree in which we observed a SNP.

7.1.3 Filtered indels and MNPs

We processed small indels and multiple-nucleotide polymorphisms (MNPs) similarly to SNPs. First, we split sites into biallelic components, then we normalized (left-aligned) representations and applied the same filters and maximum-likelihood genotyping calling approach as described for the SNPs. We then mapped 2,706 biallelic indels and MNPs to the phylogeny inferred from the SNPs and imputed genotypes accordingly.

7.2 CNV discovery and genotyping using Genome STRiP

Authors: Bob Handsaker

We performed CNV discovery and genotyping using Genome STRiP[65]. We analyzed chromosome Y in 1,234 male individuals using a pre-release version of Genome STRiP (r1.04.1447). We excluded 10 males from CNV discovery where the read depth of coverage across chromosome Y (after normalization and correction for GC-bias) was either less than 0.8 times the expected coverage or greater than 1.2 times the expected coverage (based on genome wide read depth), suggesting the potential presence of cell line specific clonal aneuploidy.

We ascertained CNVs by two methods: In the first method (discovery set 1, targeting uniquely alignable sequence) we ran the standard Genome STRiP CNV pipeline to find CNVs using read depth of coverage in uniquely alignable regions of the genome. We ran this CNV pipeline twice, once with an initial window size of 5 kb (overlapping windows by 2.5 kb) and once with an initial window size of 10 kb (overlapping windows by 5 kb). Other parameters were set to default values in each run. For both runs, the raw CNV calls were filtered using the following criteria:

- Minimum call rate: 0.8
- Minimum density of alignable positions: 0.3
- Minimum cluster separation: 5.0 (standard deviations)

In addition, for the 5 kb run, sites were excluded if they were called only in samples with high numbers of variants (more than 45 variants per sample).

We estimated the false discovery rate (FDR) for these CNV calls using the IRS method[65] and probe intensity data from Affymetrix 6.0 SNP arrays that were run on the same individuals. For sites longer then 20 kb, the estimated FDR was zero. We

included in the call set all sites longer than 20 kb and those shorter sites (under 20 kb) that contained at least one array probe and had an IRS estimated p-value<0.01.

The calls from the 5 kb and 10 kb runs were merged, re-genotyped, and duplicate calls were removed using the standard Genome STRiP duplicate removal filters. The sites were then manually reviewed and 27 calls were eliminated as being either (a) likely duplicate calls that were not detected by the default filters or (b) sites with weak evidence of copy number variation.

The second method used for CNV ascertainment (discovery set 2) targeted regions of segmental duplication. In this method, segmental duplications annotated on the UCSC genome browsers were prospectively genotyped for total copy number (using an expected reference copy number of two copies). The raw CNV calls were filtered using the following criteria (chosen based on manual review of the genotyped sites):

- Minimum call rate: 0.8
- Minimum density of alignable positions: 0.25
- Minimum cluster separation: 5.0 (standard deviations)

In the VCF file containing the CNV calls, the CNVs detected from segmental duplication analysis (discovery set 2) have site identifiers that start with "GS_SD_M2". For the segmental duplication calls, the locations of the two segmental duplication intervals are encoded in the site ID. The POS/END attributes in the VCF file specify the leftmost of these two segmental duplication intervals.

The final CNV call set consisted of 97 sites called from the first method (discovery set 1) and 13 segmental duplication sites called from the second method (discovery set 2). Copy number genotypes were encoded in the VCF file using the GT field, assuming that the reference allele has one copy for sites from discovery set 1 and that the reference allele has two copies for segmental duplication sites from discovery set 2.

7.3 Integration

The final callset with 1,233 male samples after the removal of related samples is included in the main Phase 3 release (Section 5.6). This combines the short variant (7.1) and CNV (7.2) callsets described above. The short variants were restricted to the 10.3 Mb callable region, whereas the CNV callset was not. A Y-STR callset was also produced, but not part of the official release. Details can be found in the companion paper[63].

8 Variant annotation

8.1 Functional annotation

Authors: Xiangqun Zheng-Bradley, Laura Clarke

1. VEP

Ensembl Variant Effect Predictor (VEP) (`http://www.ensembl.org/info/docs/tools/vep/index.html`) was chosen to assign functional consequences to the Phase 3 variants. VEP was run in the offline 'cache' mode, the cache files contain a dump of Ensembl gene and transcript annotations and much more. The cache files used for this release can be found at:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/source/vep_cache_ensembl_75/`

Below is an example command line for running VEP on the Phase 3 data. Options "sift" and "polyphen" were used to calculate SIFT and POLYPHEN scores. In the same process, we used `filter_vep.pl` to remove Regulatory Features annotated by VEP because subsequently, Ensembl Regulatory Build was employed to annotate Regulatory Features (see Section 8.2).

```
perl variant_effect_predictor.pl -i input.txt --format vcf \
--sift 'b' --polyphen 'b' --cache --dir \
'/path/to/the/vep/cache/files' | filter_vep.pl --format vcf \
--filter \"Feature_type != RegulatoryFeature or not Feature_type\" \
--only_matched > annotated_input.txt
```

This produced an unfiltered set of functional annotation which gives every possible consequence for each variant in the dataset. This unfiltered data can be found on the ftp site.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/unfiltered/`

2. GERP

Ensembl provided us with genome wide base by base GERP scores, which is a measurement of evolutionary conservation. For SNPs, GERP scores were added to the INFO field without additional processing. For variants with multi-nucleotide REF

alleles (deletions and other complicate events), GERP score for each base was considered and we reported the median and maximal GERP score for the span.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/source/gerp_and_multi_alignment/`

8.2 Annotating variants with the Ensembl Regulatory Build

Authors: Steven P Wilder, Nathan Johnson, Thomas Juettemann, Daniel Zerbino

The Ensembl Regulatory Build is a new process by which epigenomic data, chiefly histone marks, CTCF binding and open chromatin assays, run across multiple cell types, are processed to produce a consensus annotation of regulatory elements along the genome[66]. The output of this process is stored on the 1000 genomes ftp site:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/source/EnsemblRegBuild/`

8.2.1 Segmentation and annotation of segmentation states

We mapped ENCODE[67] and Roadmap Epigenomics[68] CTCF binding, DNase1 hypersensitivity, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K20me1 and control ChIP-seq data for 17 human cell types (A549, DND-41, GM12878, K562, H1-hESC, HepG2, HeLa-S3, HSSM, HSSMtube, HUVEC, Monocytes-CD14+, NH-A, NHDF-AD, NHEK, NHLF and Osteoblasts) using `bwa samse`[15] with default parameters.

The coverage signal of the mapped reads was merged across replicates then processed with `align2rawsignal` (`https://code.google.com/p/align2rawsignal/`), with options (`-w=180 -n=5`) We then ran a Segway[69] segmentation. The options used were (`--num-labels=25 --num-instances=10 --resolution=200 --prior-strength=1000 --ruler-scale=200 -m 1,2,3,4,5,6,7,8,9,10,11,12`). The training was run on the ENCODE pilot regions[67].

We then computed for each of the 25 states the number of cell types that have that state at every base pair of the genome. Using the overlaps of this summary function with known annotations of the genome, namely transcription start sites and exons, and experimental features namely CTCF binding sites and known chromatin repression marks, we automatically assigned preliminary labels to every state, which we

curated manually. The labels are one of: Predicted promoters, Predicted promoter flanking regions, Predicted enhancers, CTCF binding sites, Unannotated transcription factor binding sites, Unannotated open chromatin regions, Dead or Repressed.

8.2.2 Defining consensus regulatory features

To determine whether a state is useful in practice, it was compared to the overall density of transcription factor binding (as measured with ChIP-seq). Applying increasing integer cutoffs to this signal, we defined progressively smaller regions. If these regions reach 2-fold enrichment in transcription factor binding signal, then the state was retained for the build.

For every label, all the state summaries that were assigned that label and judged informative were summed into a single function. We selected the threshold that produced the highest F-score with respect to the overall TF binding signal, assuming that TF binding is a measure of regulatory activity,

To simplify visualisation and interpretation we simplified this annotation with the following rules:

- Distal enhancers that overlapped promoter-flanking regions were merged into the latter.
- Promoter flanking regions that overlapped transcription start sites were incorporated into the flanking regions of the latter features.

8.2.3 Annotating 1000 Genome Project variants

The variants, stored in VCF files were overlapped to the Ensembl Regulatory Build using BEDTools[70]. For consistency, with the VEP[71], which was also used to annotate variants in this project, a downstream awk script processed the BEDTools results to mimic the VEP output format.

8.3 Annotation of ancestral allele

Author: Javier Herrero

We derive ancestral allele (AA) using the 6-way EPO alignments available in Ensembl v71[72,73].

Note that AAs for chromosome Y were annotated in a separate process based on the inferred phylogeny (section 7).

8.3.1 SNPs

For SNPs, we use the call from EPO pipeline[74] (which infers ancestral sequences from the multiple alignment). The calls have been classified using the ancestral, the sister and the ancestral of the ancestral sequences. In this example where human, chimp, orang and macaque sequences are present only once, the sequence tree will be: (((Hsap,Ptro),Ppyg),Mmul), that is:

```
          +----- Hsap (human)
     +--(a)
     |    +--(b) Ptro (chimp)
+---(c)
|    +--------- Ppyg (orangutan)
|
+-------------- Mmul (rhesus macaque)
```

we look at the human-chimp (a) ancestral sequence and compare it to the chimp (b) and to the human-chimp-orang (c) ancestral sequences. We use the following convention for the ancestral sequence:

- uppercase when all 3 –(a), (b) and (c)– sequences agree
- lowercase when:
  - there is no ancestral sequence for the ancestral sequence, i.e. there are only two extant sequences in the alignment, but (a) and (b) agree.
  - there is a gap in the sister sequence, but (a) and (c) agree.
  - either (b) or (c) disagree with (a), but not both.
- N when both (b) and (c) disagree with (a)
- - (dash) when no there is no ancestral allele, this is a lineage-specific insertion
- . (dot) when there is no alignment, i.e. no data.

The ancestral allele can be found on the Ensembl FTP site

`ftp://ftp.ensembl.org/pub/release-74/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2`

8.3.2 INDELs

For 1bp indels, we realign both alleles to the primate sequences. We use 10bp flanking sequence each side of the indel to provide the aligner with some context, extract that section of the EPO alignment and run Ortheus[75] with each human allele alternatively. We compare the AA from both alignments and if they coincide, use this to infer the AA for that indel. In order to do make a correct inference for indels affecting a homopolymer, we consider the whole homopolymer when calling the alleles. For instance, for a T inserted next to 3 Ts we will consider that the reference allele is TTT and the alternate is TTTT. As homopolymers are not typically reflected in the VCF files, the interpretation of the reference, alternate and ancestral alleles are included in the INFO field of the final VCF file.

The call might fail for several reasons:

1. (NO_COVERAGE) The EPO alignments do not cover that area of the genome.
2. (LONG_INSERTION) There is close to the indel a long insertion in the human genome with respect to the other primate sequences, which would hamper the alignment.
3. (ALL_N) Either side of the flanking region could undetermined (i.e. all Ns in the sequence).
4. (LOW_COMPLEXITY) The indel is in a region of low complexity such that one of the two flanks is a homopolymer
5. (LOW_COMPLEXITY) The indel is in a region of low complexity like in a STR. We test this by offsetting the reference allele by 2, 3 or 4 positions, aligning it to the original sequence with MUSCLE and assessing the quality of the alignment.
6. (ALL_OTHER_Ns) After stripping all the uninformative sequences (for example, chimpanzee with only Ns in the sequence) from the alignment, we are left with the original human sequence only.

We use the same convention as before for noting the confidence in the final AA sequence.

The code used for originating these predictions is available at

`https://github.com/jherrero/ancestral-alleles/tree/v0.1-beta`

The EMF files used for the alignments can be found on the Ensembl ftp site

`ftp://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/epo_6_primate/`

9 Analysis

9.1 Imputation evaluation

Authors: Olivier Delaneau, Warren Kretzschmar, Jonathan Marchini

We carried out experiments to assess the performance of the Phase 3 haplotypes as an imputation resource. We assessed imputation accuracy using high-coverage, whole-genome sequence data made publicly available by Complete Genomics (CG).

`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130524_cgi_combined_calls/`

We used data from 59 samples that also occur in the Phase 3 haplotypes, with 10 samples from each of the populations CEU, CHS, PEL, PJL an YRI and 9 samples from LWK and filtered out all variant sites with a call rate below 90% in order to only consider very reliable sites in the analysis.

To mimic a typical imputation analysis, we created pseudo-GWAS dataset by extracting the CG SNP genotypes at all the sites included on an Illumina 1M SNP array (Human1M-Duo v3C). We then imputed all the sites not on the array using a reference panel that had the 59 samples removed. Imputation was carried out using IMPUTE2[76] which chooses a custom reference panel for each study individual in each 1 Mb segment of the genome. We set the k_{hap} parameter of IMPUTE2 to 1000. All other parameters were set to default values.

We compared the performance of the Phase 3 reference panel to impute genotypes in samples of different ancestries. For each of the 6 populations in the 59 pseudo-GWAS samples we calculated the non-reference allele frequency of each variant in the relevant continental population samples, having removed the 59 pseudo-GWAS samples. We then stratified imputed variants into allele frequency bins and calculated the squared correlation between the imputed allele dosages at variants in each bin with the masked CG genotypes. The squared correlation was then plotted against non-reference allele frequency to highlight differences between populations (Figure 4A).

All of the imputed sites in the reference panel are biallelic. However, some sites are multi-allelic sites that were decomposed into biallelic sites. We classified sites according to whether they were biallelic or multi-allelic sites, and also whether they are SNPs or Indels. In this way we were able to calculate a squared correlation analysis stratified by variants type (Extended Data Figure 9B). In this analysis the non-reference allele frequency of each variant was measured across the whole Phase 3 set of haplotypes with the 59 samples excluded.

Finally, we carried out a squared correlation analysis comparing the Phase 3 and Phase 1 haplotype panels. In each case, we imputed masked genotypes in all the 59 CG samples and across all variant types, but only at variants common to both Phase 3 and Phase 1, and matched for physical positions and alleles. In this analysis the non-reference allele frequency of each variant was measured across the whole set of Phase 3 haplotypes with the 59 samples excluded.

We created two new panels by subsetting the Phase 1 and Phase 3 panels down to their intersecting set of samples (1006 samples). We also carried out imputation experiments using these two panels. Examining imputation quality using these two new panels allowed us to compare directly the data quality between Phase 1 and Phase 3. Also, comparing the subsetted Phase 3 panel to the full Phase 3 panel allowed us to examine the role of panel size as a factor in imputation quality. Figure 4A (inset) shows a plot of the squared correlation stratified by allele frequency for all 4 of these panels. We also carried out the analysis in each of the 6 populations separately for the 4 panels (Extended Data Figure 9A).

9.2 Callable genome mask

Authors: Mary Kate Wing, Hyun Min Kang, Gonçalo R. Abecasis

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/`

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 provide an assessment of the regions of the genome that are accessible to the next-generation sequencing methods used in Phase 3, we created two genome masks.

Most project analysis did not use these hard masks for calling. Instead, the project used machine learning algorithms such as SVM 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 Phase 1 and 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.

The accessible genome masks were generated first by obtaining the per sample coverage for each reference position using bamUtil stats (`http://genome.sph.umich.edu/wiki/BamUtil:_stats`). Reads that were duplicates, QC failures, or unmapped were excluded from the coverage counts. After generating per sample coverage, we merged the samples together to generate overall per position reference coverage.

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[1]. 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 3 samples is 17,920 in autosomal chromosomes. Thus, sites with a depth of coverage of <8,960 or >35,840 were excluded. In non-pseudo-autosomal regions (non-PAR) of chrX, the depth thresholds were multiplied by a factor of 3/4 (<6,720 or >26,880). In non-PAR of chrY, the thresholds were multiplied by a factor of 1/4 (<2,240 or >8,960). Overall, this Pilot-style mask in the autosomal chromosomes results in about 6.8% of bases marked as N, 1.1% marked L, 0.2% marked H, and 2.4% marked Z. The remaining 89.4% of passed are marked passed (P) - and correspond to 95.9% 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. The average total depth of coverage across Phase 3 samples is 17,920 in the autosomal chromosomes, and the bases with lower than 0.5-fold or higher than 1.5-fold were excluded. Thus, sites with a depth of coverage of <8,960 or >26,880 were excluded for autosomal chromosomes and PAR of chrX, and the thresholds were multiplied by a factor of 3/4 (<6,720 or >20,160) for non-PAR of chrX, and by a factor of 1/4 (<2,240 or >6,720) for non-PAR of chrY.

Overall, this Strict mask in the autosomal chromosomes results in about 6.8% of bases marked N, 1.1% marked L, 0.5% marked H, 16.8% marked Z, and 3.1% marked Q. The remaining 71.7% of passed are marked passed (P) - corresponding to 76.9% of the non-N bases.

In Phase 3, 93.2% of biallelic SNPs passed SVM filtering, while 6.8% failed the SVM filtering. Only 2.8% of SNPs with a passing strict mask (P) also failed SVM filtering, while 17.0% of SNP without a passing strict mask failed SVM filtering. Looking at individual Strict masks, 10.7% of SNPs with mask L, 14.1% of SNPs with mask Q, 53.2% of SNPs with mask L, and 59.4% of SNPs with mask H failed SVM filtering.

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

9.3 Functional annotation and interpretation

Authors: Yuan Chen, Suganthi Balasubramanian, Yao Fu, Donghoon Kim, Vincenza Colonna, Heiko Horn, Jakob Berg Jespersen, Kasper Lage, Xiangqun Zheng-Bradley, Fiona Cunningham, Ian Dunham, Paul Flicek, Ekta Khurana, Daniel Zerbino, Laura Clarke, Mark Gerstein, Chris Tyler-Smith, Yali Xue

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/`

Functional annotation was carried out on the entire Phase 3 variant call set. Here, we summarise the results of this exercise, presenting the numbers per individual for different functional categories, and different frequency bins. The numbers include the number of sites, number of homozygous variants, heterozygous variants and alleles (=heterozygous variants plus 2 × homozygous variants). For the general categories described below, we counted both alternative and derived alleles. However, for GWAS variants we counted risk alle-

les, for ClinVar and HGMD-DM variants we counted disease alleles, and for indels and large deletions we counted alternative alleles. Frequency bins were singletons, singleton-0.5%, 0.5%-5% and >5% based on the global frequency. We also partitioned the genome into accessible and inaccessible regions using the Pilot mask (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.pilot_mask.autosomes.bed`).

9.3.1 Annotations and datasets

General functional categories:

VEP and Ensembl Regulatory Build annotation:

First, we identified the complete transcript for all genes using the following command:

```
zcat gencode.v19.annotation.gtf.gz | grep -v '^#' | \
awk '/\t(HAVANA|ENSEMBL)\t(CDS|start_codon|stop_codon)\t/ {print}' \
| grep -v mRNA_end_NF | grep -v mRNA_start_NF > \
gencode.v19.annotation.filtered.gtf
```

Then, we ran VEP as described in section 8.1 but with the "-pick_allele" option. This gave us a single annotation per alternative allele.

We used the same Ensembl Regulatory Build annotation as described in section 8.2.

LoF filtering by ALOFT:

Raw LoF calls were filtered to remove errors arising due to gene model annotation errors and variants unlikely to result in LoF using a procedure similar to that in the Phase 1 1000 Genomes data analysis[4]. The various LoF flags were obtained from the automated pipeline ALOFT (`https://github.com/gersteinlab/aloft`). These flags were annotated for both stop-gained and canonical splice sites.

Annotating variants with FunSeq scores

FunSeq scores[77,78] for all noncoding single nucleotide polymorphisms (SNP) in the 1000 Genomes Phase 3 dataset were calculated and stored in the VCF files. FunSeq investigates the functional impact of variants combining:

1. inter- and intra-species conservation;
2. functional genomics studies from ENCODE;

3. nucleotide-level loss- and gain-of-function events;
4. distal regulatory element-gene linkages and 5. network properties of associated genes.

A FunSeq score >1.5 was used as a cutoff for noncoding deleterious variants.

All the counts in these categories are based on the derived allele. In cases where the variant derived allele was the reference allele, we filtered them out, as all the annotation above is based on changes from reference allele to alternative allele.

Disease-related variants:

HGMD-DM entries were from Human Gene Mutation Database Professional 2014.3 version (HGMD® `http://www.hgmd.org/`).

ClinVar variants, clinvar_20141105.vcf.gz were downloaded from `ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/` in the NCBI ClinVar website. We only used sites annotated as likely pathogenic and pathogenic (CLNSIG=4 and CLNSIG=5).

GWAS catalogue variants were downloaded from the NHGRI website (http://www.genome.gov/gwastudies/). We first changed the chromosomal coordinates from GRCh38 to GRCh37, and applied the following filters, removing:

1. p-value $> 5 \times 10^{-8}$;
2. SNPs with multiple chromosomal positions;
3. Risk allele frequency unknown;
4. Risk allele is '?' or a haplotype;
5. SNPs with alleles A/T, and C/G, as we have no strand information and are not able to assign the risk allele.

This left 3,415 GWAS SNPs for further analyses.

For HGMD-DM and ClinVar variants, we based counts the on the disease allele, while the GWAS counts were based on the risk allele.

Other variants of biological interest:

Phosphorylation sites: Human phosphorylation sites were downloaded from PhosphoSitePlus®, `www.phosphosite.org` (Sep 25, 2014)[79]. Using CCDS version 15[80], the phosphorylation sites were mapped to their corresponding genomic locations. The counts are based on the derived allele.

HighD sites: For each pair of continental populations, the pairwise derived allele frequency difference (deltaDAF) was calculated for each variant. HighD sites were identified by scanning the genome using non-overlapping windows of 5,000 markers and picking the variant with the highest deltaDAF value in each window, provided that deltaDAF>0.7. HighD sites were assigned to the population with the highest DAF in the pair. The counts here are based on the derived allele. Overlap with Phase 1 was evaluated calculating the distance between Phase1 and Phase3 HighD sites using the windows of Phase 1.

9.3.2 Results

We have generated separate annotation files including all the annotations used here:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/functional_annotation/filtered/`

We have also produced a single tab-limited text file with the counts for each individual for all functional categories described here, stratified by derived allele frequency bin, genomic region accessibility, population, and super population. The files, 'functional_catagories_summary_per_individual.20150208.txt' and 'functional_categories_summary_per_superpop.20150217.txt' are available in the above ftp directory. A summary of median counts in super populations for the different categories is shown in Table 1 of the main paper and in this table:

`https://public.tableausoftware.com/views/1000G_phase3_per_individual_count_all/Sheet1?:embed=y&:showTabs=y&:display_count=no`

Here, we, highlight a few features of the median site counts.

All site counts of the general functional categories, including the variants with FunSeq score >1.5, showed a similar trend with the order EAS>AFR>SAS>AMR>EUR for singleton variants, the order of AFR≫SAS>EAS>AMR>EUR for rare variants (singleton-0.5%), AFR≫AMR>EUR>SAS>EAS for relatively common variants (0.5%-5%) and for the very common variants, the counts are not much different among these super populations except EAS were lower. LoF variants (stop gained, splice acceptor, splice donor and frameshift) showed very similar number across different populations for singletons (frameshift not applicable here) and common ones (>5%), but much higher numbers in AFR for doubletons to 5%. However, all the trends evened out when we corrected them by heterozygosity. Nevertheless, the raw counts illustrate the general population diversity features of the different super populations.

For HGMD-DM variants, we saw more variants per individual in the EUR for the very rare ones (singletons-0.5%) which are most likely to be truly pathogenic, but similar numbers in the different populations for the common variants (>0.5%). However, for ClinVar variants, we saw very similar numbers per super population in different frequency bins, and the common ones are much higher than for HGMD-DM variants, which suggests different annotation criteria in the different databases and raises questions about the true clinical significance of the common ClinVar variants. As we expect, we only see GWAS sites in the common frequency bins (>5%) in all super populations, and the numbers are very similar in all the super populations, except the EAS are a little lower.

In total, 5,159 phosphorylation sites overlap with 1000 Genomes Phase 3 variants. Phosphorylation sites observed in more than five high-throughput experiments were significantly depleted for non-synonymous coding changes (odds-ratio = 0.81, $P = 1.7 \times 10^{-3}$ by Fisher's exact test). This trend fits the hypothesis that sites that are more often observed are more likely to be functional (and thus less likely to be mutated). We did not see much difference among the super populations and frequency bins.

Derived allele frequencies (DAFs) at SNVs were compared between pairs of super populations as described previously[4,81]. We identified 498 unique sites with extreme DAF difference (HighD sites). About 80% of them are located in transcripts (mostly coding), especially within introns. One third of the sites identified in Phase3 exactly overlap with Phase 1. More generally, 44% of Phase 3 HighD sites are ≤6 kb from Phase 1 HighD sites. The remaining 56% are found at a median distance of 43kb. This discordance reflects the different sets of samples used to define the super populations and the higher number of markers in Phase 3, which leads to different windows being defined.

HighD sites are enriched for positive selection and among those we identified are, as expected, four missense variants known to be under positive selection (rs1426654 in SLC24A5, rs3827760 in EDAR, rs16891982 in SLC45A2 and rs1871534 in SLC39A4 (ZIP4)), and rs2814778 in the 5' untranslated region of DARC. Among novel examples of HighD sites we note another example of differentiation between EAS and SAS, the intronic variant rs200189385 of the Rho GTPase activating protein 42 (ARHGAP42; DAF=0.79 in EAS and 0.02 in SAS) expressed specifically in smooth muscle cells and implicated in the regulation of the vascular tone[82].

Counts per individual are highest in EUR, intermediate in AFR and EAS, and lowest in AMR and SAS. The low number in AMR is expected because of the extensive European and African admixture in these samples. The low number in SAS points to a lack of extreme drift and positive selection detectable by this approach in this

region.

9.4 Estimating effective population size with low coverage data

Authors: Anthony Marcketta and Adam Auton

The pairwise sequentially Markovian coalescent (PSMC) model[83] is a method for estimating population size from genetic sequence data. This program uses the distribution of heterozygotes throughout the genome to determine an estimate of the time to the most recent common ancestor of a segment of sequence.

In past applications of PSMC, the method has been applied to high coverage sequence data. To apply PSMC to the 1000 Genomes callset, we first converted the variants from the VCF files into PSMC input using 100 base pair bins. Bins were encoded as follows:

- Bins were labelled as **Missing** if >= 90 bases were not defined in the reference genome.
- Bins were labelled as **Heterozygous** if >= 10 bases were defined in the reference genome, and >= 1 heterozygote site was observed.
- Bins were labelled as **Homozygous** if >= 10 bases were defined in the reference genome, and no heterozygote sites were observed.

To ensure that the analysis was not overly affected by variations in sequence coverage, we applied a mask to ensure that only well called regions of the genome were included. Specifically, we applied a mask based on the union of the 1000 Genomes 'strict' mask and the negation of the build 37 'low complexity regions' mask. Any bases outside of the mask were considered to be missing data. These masks can be found on the 1000 Genomes FTP at the following locations:

- `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/low_complexity_regions//hs37d5-LCRs.20140224.bed.gz`
- `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20140520.strict_mask.autosomes.bed`

PSMC was run on each of 2,504 individuals included in the 1000 Genomes Phase 3 VCF files. For plotting, individuals were grouped by their populations, the median effective population size within each segment was determined. The resulting lines were smoothed by fitting a cubic spline and using the PSMC bin midpoints.

To determine if PSMC could properly estimate population sizes using imputed data, we also applied PSMC to the 26 high coverage PCR-free sequenced individuals for comparison to the low coverage results. BAM files were downloaded from the 1000 Genomes FTP site and were converted into genome-wide consensus FASTQ files using the samtools mpileup function and the vcfutils vcf2fq function. These FASTQ files were also masked by replacing any bases outside the previously mentioned 'combined mask' as missing data. They were then pared down into 100 base pair bins as previously described and processed using PSMC. The results were plotted against the corresponding low coverage estimates (Extended Data Figure 7). The results are largely similar, although the low coverage estimates slightly underestimating recent population expansions. This is most likely due to limited power of the low coverage sequencing to detect particularly rare variation.

PSMC output for the above analysis can be found at the following location:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/psmc/`

9.5 Identity by Descent (IBD) segment sharing within and between populations

Authors: Sharon R. Browning and Brian L. Browning

To infer regions of IBD within the 1000 Genomes Phase 3 data, we first filtered the phased genotype data to retain only diallelic SNVs with more than ten copies of the minor allele. We inferred IBD segments using Refined IBD (Beagle v4)[84]. The centiMorgan (cM) length of each IBD segment and chromosome were obtained from the Hapmap genetic map[27]. We filled gaps between adjacent IBD segments if the gaps contained a low rate of homozyogous discordances, because phasing and genotype errors can create apparent gaps in longer IBD segments. We retained only segments of final length greater than 5 cM.

We calculated a kinship coefficient for each pair of individuals by summing the lengths of the autosomal IBD segments and dividing by four times the length of the autosomal genome. We excluded 82 pairs of individuals with an estimated kinship coefficient greater than 0.05. Only one of the excluded pairs crosses a population boundary (an ITU/STU pair). We calculated average kinship coefficients within each population

and between each pair of populations by averaging the kinship coefficients for each non-excluded pair of individuals.

We also calculated average kinship coefficients within and between populations from the IBD segments on the X chromosome. We summed the lengths of the X chromosome IBD segments in each non-excluded pair of individuals and divided by the X chromosome length and by the number of analysed haplotype pairs (4 haplotype pairs per female-female individual pair, 2 haplotype pairs per female-male pair, and 1 haplotype pair per male-male pair).

Estimates of IBD sharing between pairs of individuals can be found at the following location:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/ibd_by_pair/`

9.6 Multipopulation eQTL analysis

Authors: Marianne DeGorter and Stephen Montgomery

Gene expression was quantified using the Human-6 Expression BeadChip v2 from Illumina (San Diego, California, USA), and normalized as previously described[85]. Normalized expression values for 21,800 probes corresponding to 18,226 autosomal genes were used for eQTL discovery.

Sixty-nine samples from populations in Phase 3 for which gene expression data was available (CEU, CHB, GIH, JPT, LWK and YRI) were used. Variant sites were filtered for a minor allele frequency of at least 1% across all six populations, and required to have at least three alleles in the population to be tested for association to gene expression.

For each population, eQTLs were assessed using Matrix eQTL version 2.1.0[86] using linear regression. Tested variants were within 1 Mb of the transcription start site. Genotypes were coded as 0, 1 or 2 alternate alleles. eQTL p-values were Bonferroni adjusted per gene in order to choose the best variant or linked set of variants for each gene. Bonferroni adjusted p-values for the best loci per gene were then used to calculate the Benjamini-Hochberg false discovery rate (FDR) of 0.05.

For multi-population eQTL meta-analysis, eQTL discovered at FDR<0.05 defined the eQTL gene list within each population. In the remaining five populations, nominal p-values for variants within 100,000 bp of the best variant in the discovery population were calculated using the Fisher's combined probability test. Among variants in LD ($r^2 > 0.6$) with the best variant in the discovery population, the

lowest p-values was used to choose a best meta-analysis variant for each eQTL gene. Variants were intersected with transcription factor binding sites (TFBS) in LCLs identified by the ENCODE project (Version 3)[67]. Enrichment in TFBS before and after fine-mapping was assessed by McNemar's test.

To check for the possible effects of variants located within probes and now identified with whole genome sequencing data, we retested eQTL discovery adjusting those probes which overlapped variants using linear regression. Here, probe annotation was provided by reMOAT[87]. This resulted in normalized expression values for 19,517 probes corresponding to 16,122 autosomal. Results from eQTL analysis remained robust to exclusion of probes containing potential variants. Files from both eQTL analyses are available on the ftp site.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/eqtl_analysis/`

9.7 Assessment of population structure using ADMIXTURE

Authors: Anthony Marcketta and Adam Auton

In order to investigate population structure within the Phase 3 dataset, we performed an unsupervised clustering analysis using the ADMIXTURE program[88]. Despite being more efficient than many similar algorithms, it is computationally prohibitive to apply structure to the entire 1000 Genomes dataset. To reduce the dataset, we used VCFtools to only keep biallelic, non-singleton SNV sites that were a minimum of 2 KB apart from each other. These variants were then merged and converted into PLINK format, giving us 1,286,213 throughout all the autosomes. PLINK was then used to filter the variants further, keeping only variants with a minor allele frequency of at least 0.05, and to convert the variants into binary PED format. After this filter 193,634 sites remained with data from 2,504 individuals in 26 populations.

We ran the ADMIXTURE software using default parameters with K values ranging from 5 through 12. Individuals within the output was sorted by the dominant cluster within each population, and plotted using a modified version of the 'distruct.py' script from the fastSTRUCTURE program[89]. The results are shown in Extended Data Figure 5.

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/admixture_`

files/

9.8 Estimating the age of f_2 variants

Author: Iain Mathieson

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/f2_analysis/

We estimated the age of f_2 variants as described by Mathieson and McVean[90]. Briefly, we defined f_2 variants to be SNPs where the majority allele appears exactly twice in the dataset, in different individuals. We first identified all the f_2 variants in the dataset. Each of these identifies a position where two samples are genealogical nearest neighbours and therefore share a recent haplotype. Note that the age of the haplotype provides a lower bound for the age of the variant. We estimated the length of these haplotypes by scanning left and right across the genome from each f_2 variant until we found sites inconsistent with a shared haplotype. 8,544,594 f_2 variants produced 5,284,225 f_2 haplotypes since some haplotypes contain more than one variant. We estimated the genetic length of the haplotype using the combined HapMap recombination map[27], and calculated the maximum likelihood estimate of the age of each haplotype given its genetic length, an estimate of the uncertainty in its length (computed from the data), and the number of singletons on the haplotype. We made two changes compared to the originally reported analysis of the Phase 1 data[90]. First, we used the sequence data for all analyses (instead of combining the sequence and array genotype data as in the original analysis). Second, we estimated a power to detect singletons of 50% rather than 33%.

9.9 Variant detection sensitivity and genotype accuracy

Authors: Hyun Min Kang, Goo Jun

We evaluated the variant detection sensitivity and genotype accuracy by comparing with variant calls obtained from deeply sequenced complete genomics samples described in Section 3.5.3. Variants with call rate less than 90% or inbreeding coefficient less than -0.1 (showing excessive heterozygosity) were further filtered out before the evaluation. All variants are also normalized in left-aligned parsimonious form using 'vt normalize' tool available at http://genome.sph.umich.edu/wiki/vt, and redundant variants after normalization were removed. The filtered and normalized calls are available at:

ftp://ftp.1000genomes.ebi.ac.uk:/vol1/ftp/release/20130502/supporting/cgi_variant_calls/filtered_calls

The variant detection sensitivity was obtained by calculating the fraction of CG variants overlapping with 1000G variants, among the variants polymorphic within the 284 overlapping samples, based on exact matches of positions and alleles. Compared to the main callset, Complete Genomic indels were slightly rarer in low-complexity sequences and slightly shorter, and hence may provide an imperfect assessment of indel sensitivity. For large deletions, we used reciprocal overlap (RO) threshold of 0.8 to determine overlap between variants. For SNPs, we also estimated SNP detection sensitivity from the Haplotype Reference Consortium (HRC) variants (excluding 1000 Genomes samples), by calculating the fraction of overlapping variants among the HRC variants at each frequency bin (Extended Data Figure 2A). We also compared the variant detection sensitivity between Phase 3 and Phase 1 by focusing on 170 samples sequenced in both releases (Extended Data Figure 2B).

We also evaluated the genotype accuracy across different variant types by calculating heterozygous genotype discordance within the overlapping variants between Phase 3 and CG genotypes (Extended Data Figure 2C). We observed that SNPs and bi-allelic variants have higher genotype accuracy than indels and multi-allelic variants, respectively. When we compared to the genotype accuracy between Phase 3 and Phase 1 (Extended Data Figure 2D), we observed that the heterozygous genotype discordance was reduced by 62% for SNPs and 72% for indels (Supplementary Information Table 4).

We also evaluated the SNP detection sensitivity by calculating the fraction of overlapping SNPs between Phase 3 and CG among the variants carrying non-reference genotypes for each of 284 CG sample. The heterozygous SNP genotype accuracy was also calculated within the overlapping variants per each sample. The variant detection sensitivity and genotype accuracy highly depends on the sequencing depth. For example, when sequencing depth increased from 5× to 10×, the average variant detection sensitivity increased from 98.3% to 98.6%, and the heterozygous SNP genotype discordance was reduced from 0.61% to 0.18% (Extended Data Figure 2E and 2F).

9.10 Genotype covariance

Author: Shane McCarthy

Genotype covariance between samples x and y in the upper diagonal of Extended Data Figure 6A was calculated as:

$$C(x,y) = \sum_i \frac{(x_i - 2f_i)(y_i - 2f_i)}{2f_i(1-f_i)}$$

where the sum is over all sites, x_i is the dosage at site i in sample x, y_i is the dosage at site i in sample y, and f_i is the allele frequency of site i. Only biallelic autosomal SNPs were included. Code is available as a plugin for bcftools (`https://github.com/samtools/bcftools`) here:

`https://github.com/mcshane/bcftools_plugins/blob/a990215b4a762d1a9e95259b8adfc/covariance.c`

9.11 Estimating GWAS Type 1 error rate

Author: Lars G. Fritsche, Hyun Min Kang, Gonçalo R. Abecasis

To evaluate the impact of our new reference panel on GWAS, we re-analyzed the Michigan Mayo AREDS Penn (MMAP) study of age-related macular degeneration (AMD) genomewide association study totaling 2,136 cases and 1,139 controls of European/Caucasion ancestry[91].

We imputed the pre-phased MMAP GWAS data using Minimac 3 (`http://genome.sph.umich.edu/wiki/Minimac3`) and the following three reference panels: 1000 Genome Project Phase 3 Release 5 (ALL, $N = 2,504$), 1000 Genome Project Phase 1 Release 3 (ALL, $N = 1,092$), and HapMap 2 Release 22 (CEU, $N = 120$). After imputation, we excluded monomorphic and poorly imputed variants ($R^2 < 0.3$). We analysed the GWAS using a logistic regression score test implemented in EPACTS (`http://genome.sph.umich.edu/wiki/EPACTS`) and estimated a Bayesian Credible set of likely functional variants using the approach of Maller *et al.*[92] (Supplementary Information Table 9). Two principal components were included as covariates in the analysis and the genomic control value was <1.0× in each analysis, indicating type I error rates were well controlled in all analyses.

To empirically determine the type 1 error rate we permuted the case-control status 1,000 times and re-analysed the GWAS data using the logistic regression score test implemented in EPACTS. We noted the smallest p-values of each permuted GWAS and estimated the 5% significance level for a GWAS as the 50^{th} smallest p-value. We estimated the effective number of independent tests by calculating the number of variants that would match this empirical threshold using a Bonferroni correction. This estimate depends on the number of variants that can be imputed, on the redundancy or linkage disequilibrium between them, and on genotyping error (when genotyping error is higher or imputation quality is lower, redundancy between otherwise similar variants decreases and the number of independent tests increases). Our results indicate that the commonly used significance threshold for genome wide significance of 5×10^{-8} (Bonferroni correction for 1,000,000 independent test) is

appropriate for HapMap2 imputed data (95% of observed top GWAS signals with p-values $\geq 6.23 \times 10^{-8}$; or ~800,000 independent tests) but that more stringent thresholds in the range of 1×10^{-8} and 1.5×10^{-8} are required for 1,000 Genomes Project Phase 1 and Phase 3 imputed data sets (95% of observed top GWAS signals with p-values $\geq 1.31 \times 10^{-8}$ and $\geq 1.47 \times 10^{-8}$; corresponding ~3.4–3.8M independent tests, respectively).

10 Accessing 1000 Genomes data

Authors: Laura Clarke, Xiangqun Zheng-Bradley

A full description of data management and community access can be found in Clarke *et al.*[93]. 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/`

The FTP structure places different data types in different directories. The Phase 3 sequence and alignment files are located under `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data`. There are index files which list all the sequence and alignment files which were used for Phase 3.

- Sequence Index including SRA and ENA accessions for all archived data: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.analysis.sequence.index`
- Low Coverage Alignment Index: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.low_coverage.alignment.index`
- Exome Alignment Index: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.exome.alignment.index`

The variants identified as part of the Phase 3 analysis have all been submitted to dbSNP or the DGVa as different the variant classes required. Our own VCF files with genotypes for each individual are available from our Phase 3 release directory.

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

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.

10.1 GRCh38 resources

As much of the community is moving toward the new version of the human reference (GRCh38) we also are building resources for the new genome.

dbSNP have remapped the 1000 Genomes variants as part of their release - v142 for the autosomes and v143 for the sex chromosomes. The dbSNP remap for the 1000 Genomes sites is available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/

We are also in the process of remapping the 1000 Genomes sequence data to the new assembly. We hope to release the low coverage alignments in late summer 2015. We are using the alt aware version of bwa mem to align the data to take advantage of the large number of alternative loci present in GRCh38. The genome and mapping resources we are using for this process can be found in ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/.

11 References

1. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

2. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

3. The H3Africa Consortium. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

4. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

5. Freshney, R. I. & Freshney, M. G. *Culture of immortalized cells* (Wiley-Liss, 1996).

6. Coriell Institute. Frequently Asked Questions about Lymphoblastoid Cell Cultures (2012). URL `http://ccr.coriell.org/Sections/Support/Global/Lymphoblastoid.aspx?PgId=213`.

7. Coriell Institute. Genotyping with Microsatellites Assures Cell Line Identity and Culture Purity (2012). URL `http://ccr.coriell.org/Sections/Support/Global/QCgenotype.aspx?PgId=412`.

8. Stevens, E. L. *et al.* Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genetics* **7**, e1002287 (2011).

9. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* **12**, R1 (2011).

10. Baylor College of Medicine, Human Genome Sequencing Center. Illumina Barcoded Paired-End Capture Library Preparation URL `https://hgsc.bcm.edu/sites/default/files/documents/Illumina_Barcoded_Paired-End_Capture_Library_Preparation.pdf`.

11. Illumina. TruSeq® DNA PCR-Free Sample Preparation Guide (2013). URL `http://supportres.illumina.com/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdnapcrfree/truseq-dna-pcr-free-sample-prep-guide-15036187-b.pdf`.

12. Illumina. HiSeq® 2000 System User Guide (2013). URL `http://supportres.illumina.com/documents/documentation/system_documentation/hiseq2000/hiseq2000_ug_15011190_r.pdf`.

13. Affymetrix. Affymetrix Genome-Wide Human SNP Nsp/Sty 6.0 User Guide, Rev 4 (2008).

14. Li, H. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* btu356 (2014).

15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (2009).

16. Li, H. *et al.* The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

17. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).

18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv preprint* **arXiv:1303.3997** (2013).

19. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *The American Journal of Human Genetics* **91**, 839–848 (2012).

20. Wang, Y., Lu, J., Yu, J., Gibbs, R. A. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Research* **23**, 833–842 (2013).

21. Challis, D. *et al.* An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* **13**, 8 (2012).

22. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research* **20**, 273–280 (2010).

23. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv preprint* **arXiv:1207.3907** (2012).

24. Durbin, R., Eddy, S. R., Krogh, A. & Mitchinson, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. IBSN: 978-0521629713 (Cambridge University Press, 1998).

25. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

26. DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics* **43**, 491–498 (2011).

27. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 851–861 (2007).

28. Rimmer, A. *et al.* Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**, 912–918 (2014).

29. Albers, C. *et al.* Dindel: Accurate indel calls from short-read data. *Genome Research* **21**, 961–973 (2011).

30. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics* **44**, 226–232 (2012).

31. Simpson, J. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).

32. Simpson, J. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**, 549–556 (2012).

33. Simpson, J. T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* btu023 (2014).

34. Cleary, J. G. *et al.* Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *Journal of Computational Biology* **21**, 405–419 (2014).

35. Shringarpure, S. S., Carroll, A., De La Vega, F. M. & Bustamante, C. D. Inexpensive and highly reproducible cloud-based variant calling of 2,535 human genomes. *PLoS One* **10**, e0129277 (2015).

36. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research* **22**, 1154–1162 (2012).

37. Willems, T. *et al.* The landscape of human STR variation. *Genome Research* **24**, 1894–904 (2014).

38. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**, e70 (2005).

39. Sudmant, P. H. *et al.* A haplotype resolved map of structural variation in 2,504 human genomes **(submitted)**.

40. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).

41. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

42. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19**, 1270–1278 (2009).

43. 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 (2011).

44. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods* **7**, 576–577 (2010).

45. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).

46. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics* **43**, 269–276 (2011).

47. 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 (2009).

48. Gardner, E. J. & Devine, S. E. MELT: Mobile Element Location Tool **(in preparation)**.

49. Dayama, G., Emery, S. B., Kidd, J. M. & Mills, R. E. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research* **42**, 12640–12649 (2014).

50. Delaneau, O., Marchini, J. & The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5** (2014).

51. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).

52. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).

53. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91 (2013).

54. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

56. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).

57. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–569 (2013).

58. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* (2014).

59. Parsons, J. Miropeats: graphical DNA sequence comparisons. *Computer applications in the biosciences: CABIOS* **11**, 615–619 (1995).

60. Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research* **24**, 2066–2076 (2014).

61. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnology* **29**, 59–63 (2011).

62. Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Research* gkr1042 (2011).

63. The 1000 Genomes Project Chromosome Y Group. Punctuated bursts in human male demography from analysis of 1,244 worldwide Y-chromosomal sequences **(in preparation)**.

64. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).

65. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature Genetics* **47**, 296–303 (2015).

66. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biology* **16** (2015).

67. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

68. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology* **28**, 1045–1048 (2010).

69. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012).

70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

71. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).

72. Beal, K., Flicek, P. & Herrero, J. Inference of the ancestral allele for any 1-bp variant in the human genome **(in preparation)**.

73. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Research* gkt1196 (2013).

74. Beal, K. *et al.* Ensembl Comparative Genomics Resources **(in preparation)**.

75. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research* **18**, 1829–1843 (2008).

76. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics* **1**, 457–470 (2011).

77. Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology* **15**, 480 (2014).

78. Khurana, E. *et al.* Integrative annotation of variants from 1,092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).

79. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research* gkr1122 (2011).

80. Farrell, C. M. *et al.* Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Research* **42**, D865–D872 (2014).

81. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology* **15**, R88 (2014).

82. Bai, X. *et al.* The smooth muscle-selective RhoGAP GRAF3 is a critical regulator of vascular tone and hypertension. *Nature Communications* **4** (2013).

83. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

84. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

85. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics* **8**, e1002639 (2012).

86. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

87. Barbosa-Morais, N. L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Research* **38**, e17–e17 (2010).

88. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).

89. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

90. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genetics* **10**, e1004528 (2014).

91. Chen, W. *et al.* Genetic variants near TIMP3 and high-density lipoprotein–associated loci influence susceptibility to age-related macular degeneration. *Proceedings of the National Academy of Sciences* **107**, 7401–7406 (2010).

92. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics* **44**, 1294–1301 (2012).

93. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nature Methods* **9**, 459–462 (2012).

Supplementary Information Tables

Population		Code	Population Color	Continental Group Color	Analysis Panel	Phase 1	Phase 3
African ancestry							
Esan in Nigeria	Esan	ESN			AFR		99
Gambian in Western Division, Mandinka	Gambian	GWD			AFR		113
Luhya in Webuye, Kenya	Luhya	LWK			AFR	97	99
Mende in Sierra Leone	Mende	MSL			AFR		85
Yoruba in Ibadan, Nigeria	Yoruba	YRI			AFR	88	108
African Caribbean in Barbados	Barbadian	ACB			AFR/AMR		96
People with African Ancestry in Southwest USA	African-American SW	ASW			AFR/AMR	61	61
Americas							
Colombians in Medellin, Colombia	Colombian	CLM			AMR	60	94
People with Mexican Ancestry in Los Angeles, CA, USA	Mexican-American	MXL			AMR	66	64
Peruvians in Lima, Peru	Peruvian	PEL			AMR		85
Puerto Ricans in Puerto Rico	Puerto Rican	PUR			AMR	55	104
East Asian ancestry							
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX			EAS		93
Han Chinese in Beijing, China	Han Chinese	CHB			EAS	97	103
Southern Han Chinese	Southern Han Chinese	CHS			EAS	100	105
Japanese in Tokyo, Japan	Japanese	JPT			EAS	89	104
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV			EAS		99
European ancestry							
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU			EUR	85	99
British in England and Scotland	British	GBR			EUR	89	91
Finnish in Finland	Finnish	FIN			EUR	93	99
Iberian Populations in Spain	Spanish	IBS			EUR	14	107
Toscani in Italia	Tuscan	TSI			EUR	98	107
South Asian ancestry							
Bengali in Bangladesh	Bengali	BEB			SAS		86
Gujarati Indians in Houston, TX, USA	Gujarati	GIH			SAS		103
Indian Telugu in the UK	Telugu	ITU			SAS		102
Punjabi in Lahore, Pakistan	Punjabi	PJL			SAS		96
Sri Lankan Tamil in the UK	Tamil	STU			SAS		102
Total						**1092**	**2504**

Supplementary Information Table 1: Population names and abbreviations.

Callset	Method	SNPs	Indels	SNP Ts/Tv (*)	% of CG SNPs (*,+)	% CG /genome (*,+)	SNP FDR (*,#)
BI1	UnifiedGenotyper	77.4m	4.16m	2.08	87.9%	99.0%	2.1%
BCM	SNPTools	73.0m	-	1.96	83.5%	99.2%	3.0%
UM	GotCloud	69.1m	2.85m	2.15	85.5%	98.7%	1.0%
BC	FreeBayes	66.1m	1.98m	2.17	83.1%	98.6%	0.8%
STN	RTG-Variant	56.5m	5.37m	2.06	76.5%	97.2%	0.1%
OX1	Platypus	56.3m	5.36m	2.18	25.6%	58.2%	0.6%
SI1	samtools	49.9m	4.42m	2.21	77.8%	98.1%	0.6%
BI2	HaplotypeCaller	29.6m	2.39m	2.17	63.4%	95.5%	1.5%
OX2	Cortex	14.5m	1.05m	1.98	81.3%	98.2%	0.1%
SI2	SGA	13.5m	1.37m	2.14	39.6%	77.9%	0.9%
Union	Simple Union	91.9m	9.23m	1.86	89.3%	99.5%	3.5%
Phase 3	**Integration**	**81.4m**	**3.37m**	**2.09**	**88.2%**	**99.1%**	**2.1%**

Supplementary Information Table 2: SNP and Indel input callset details. * Biallelic SNPs, + Percentage of CG variants rediscovered by callset, # Maximum likelihood estimate of the FDR based on validation in 26 high coverage PCR-free samples.

	Autosomes	Exome target regions**	chrX***	chrY***	Totals
Samples	2,504	2,504	2,504	1,233	-
Total Raw Bases (Gb)	85,426	18,273	3,213	291	-
Mean Mapped Depth (X)*	8.45	75.25	6.20	2.60	-
Total Variant Sites	84,801,880	1,416,049	3,468,093	62,042	88,332,015
Biallelic SNPs	81,102,777	1,383,927	3,223,927	60,505	84,387,209
Indels	3,196,364	19,832	212,196	1,427	3,409,987
Mean Indel Length (bp)	2.94	3.46	2.64	2.00	-
Multiallelic sites	444,026	6,153	30,996	-	475,022
Multiallelic SNPs	274,425	4,706	15,055	-	289,480
Multiallelic Indels	169,601	1,447	15,941	-	185,542
Structural Variants	58,713	6,137	974	110	59,797
ALU Insertion	12,491	52	-	-	12,491
LINE1 Insertion	2,910	10	-	-	2,910
Large Deletion	33,336	2,684	974	-	34,310
Duplication	5,896	2,513	-	-	5,896
SVA Insertion	822	5	-	-	822
Other Insertion	165	1	-	-	165
Inversion	100	8	-	-	100
CNV	2,993	864	-	110	3,103

Supplementary Information Table 3: Integrated callset summary. *Assuming 2.84Gb as the genome size. The mapping of exome sequence to targeted pull down regions was calculated by Picard function *calculateHsMetrics*. **The exome targeted regions were exome pulldown targets derived from CCDS (NimbleGen EZ Exome v1 and Agilent SureSelect v2). These variant totals are included in the other columns. ***chrX and chrY statistics are for the entire chromosomes.

Variant Type	Phase	All	NonRef	Heterozygotes
SNPs	Phase 3	0.06%	0.57%	0.47%
SNPs	Phase 1	0.19%	1.71%	1.23%
Indels	Phase 3	0.16%	1.01%	0.62%
Indels	Phase 1	0.52%	3.30%	2.19%

Supplementary Information Table 4: Genotype discordance with Complete Genomics data. Genotype concordance between the 1000 Genomes callset and the Complete Genomics dataset was calculated using 170 overlapping samples at 19.3M overlapping SNPs, and 493k overlapping indels. *All* - All sites overlapping between the project calls and called Complete Genomics sites. *NonRef* - Sites at which either the project calls or Complete Genomics called a non-reference allele. *Heterozygotes* - Sites at which Complete Genomics called a heterozygous genotype.

doi:10.1038/nature15393

Population	CDX	KHV	CHS	CHB	JPT	PEL	MXL	CLM	PUR	IBS	TSI	CEU	GBR	FIN	GIH	PJL	ITU	STU	BEB	ACB	ASW	LWK	ESN	YRI	MSL	GWD
CDX	0	0.0019	0.0049	0.0086	0.0169	0.0864	0.0708	0.0766	0.0826	0.105	0.1051	0.106	0.1062	0.0991	0.073	0.0705	0.0685	0.0671	0.056	0.1405	0.1254	0.1524	0.1631	0.1613	0.1636	0.1581
KHV	0.0019	0	0.0033	0.0062	0.014	0.0823	0.0665	0.0726	0.0788	0.1008	0.1009	0.1016	0.1018	0.095	0.0687	0.0664	0.0643	0.0629	0.0521	0.1379	0.1224	0.1499	0.1608	0.1589	0.1611	0.1557
CHS	0.0049	0.0033	0	0.0011	0.0088	0.082	0.0677	0.0749	0.0816	0.1042	0.1043	0.105	0.1053	0.0977	0.0726	0.0702	0.0682	0.0669	0.0558	0.1418	0.1263	0.1538	0.1647	0.1627	0.1651	0.1595
CHB	0.0086	0.0062	0.0011	0	0.0069	0.0795	0.0652	0.0727	0.0796	0.1022	0.1023	0.1029	0.1032	0.0956	0.0712	0.0687	0.067	0.0657	0.0547	0.1405	0.1248	0.1525	0.1634	0.1614	0.1638	0.1582
JPT	0.0169	0.014	0.0088	0.0069	0	0.0798	0.0659	0.0738	0.0809	0.1039	0.1041	0.1046	0.1049	0.0971	0.0724	0.0701	0.0683	0.0668	0.0564	0.1419	0.1262	0.1538	0.1648	0.1628	0.1652	0.1595
PEL	0.0864	0.0823	0.082	0.0795	0.0798	0	0.0174	0.0391	0.0564	0.0856	0.0865	0.0844	0.0848	0.0806	0.074	0.0705	0.0742	0.0739	0.068	0.1377	0.119	0.1512	0.1619	0.1599	0.162	0.1567
MXL	0.0708	0.0665	0.0677	0.0652	0.0659	0.0174	0	0.0093	0.0189	0.0359	0.0367	0.0358	0.036	0.0356	0.038	0.0337	0.0398	0.0398	0.0353	0.1033	0.0829	0.1178	0.1289	0.1272	0.1282	0.1238
CLM	0.0766	0.0726	0.0749	0.0727	0.0738	0.0391	0.0093	0	0.0056	0.014	0.0152	0.0152	0.0154	0.0184	0.0281	0.0234	0.0311	0.0312	0.0284	0.0879	0.0677	0.103	0.114	0.1125	0.1135	0.1091
PUR	0.0826	0.0788	0.0816	0.0796	0.0809	0.0564	0.0189	0.0056	0	0.0087	0.0097	0.0108	0.0109	0.0155	0.0275	0.0226	0.0309	0.0311	0.0292	0.0756	0.0566	0.0905	0.101	0.0995	0.1004	0.0962
IBS	0.105	0.1008	0.1042	0.1022	0.1039	0.0856	0.0359	0.014	0.0087	0	0.0016	0.0024	0.0024	0.0103	0.0343	0.0286	0.0395	0.0401	0.0393	0.1065	0.0861	0.1232	0.1354	0.1336	0.1352	0.1295
TSI	0.1051	0.1009	0.1043	0.1023	0.1041	0.0865	0.0367	0.0152	0.0097	0.0016	0	0.0036	0.0038	0.0118	0.0329	0.0273	0.0381	0.0388	0.0382	0.1078	0.0875	0.1244	0.1367	0.1349	0.1365	0.1308
CEU	0.106	0.1016	0.105	0.1029	0.1046	0.0844	0.0358	0.0152	0.0108	0.0024	0.0036	0	0.0003	0.0064	0.0338	0.0281	0.0396	0.0401	0.039	0.1107	0.0899	0.1278	0.14	0.1383	0.1398	0.1342
GBR	0.1062	0.1018	0.1053	0.1032	0.1049	0.0848	0.036	0.0154	0.0109	0.0024	0.0038	0.0003	0	0.0068	0.0341	0.0284	0.0399	0.0405	0.0393	0.1106	0.0899	0.1276	0.1398	0.1381	0.1396	0.1341
FIN	0.0991	0.095	0.0977	0.0956	0.0971	0.0806	0.0356	0.0184	0.0155	0.0103	0.0118	0.0064	0.0068	0	0.035	0.0296	0.0402	0.0406	0.0386	0.1135	0.0928	0.1304	0.1425	0.1407	0.1423	0.1368
GIH	0.073	0.0687	0.0726	0.0712	0.0724	0.074	0.038	0.0281	0.0275	0.0343	0.0329	0.0338	0.0341	0.035	0	0.0037	0.0039	0.0044	0.0045	0.1028	0.0841	0.1169	0.1287	0.1271	0.1286	0.1236
PJL	0.0705	0.0664	0.0702	0.0687	0.0701	0.0705	0.0337	0.0234	0.0226	0.0286	0.0273	0.0281	0.0284	0.0296	0.0037	0	0.0033	0.0037	0.0038	0.0991	0.0802	0.1135	0.1251	0.1237	0.1249	0.1201
ITU	0.0685	0.0643	0.0682	0.067	0.0683	0.0742	0.0398	0.0311	0.0309	0.0395	0.0381	0.0396	0.0399	0.0402	0.0039	0.0033	0	0.0012	0.0024	0.1021	0.0839	0.1158	0.1274	0.1259	0.1272	0.1224
STU	0.0671	0.0629	0.0669	0.0657	0.0668	0.0739	0.0398	0.0312	0.0311	0.0401	0.0388	0.0401	0.0405	0.0406	0.0044	0.0037	0.0012	0	0.0022	0.1016	0.0835	0.1152	0.1268	0.1252	0.1265	0.1217
BEB	0.056	0.0521	0.0558	0.0547	0.0564	0.068	0.0353	0.0284	0.0292	0.0393	0.0382	0.039	0.0393	0.0386	0.0045	0.0038	0.0024	0.0022	0	0.1001	0.0819	0.1138	0.1252	0.1237	0.1249	0.1204
ACB	0.1405	0.1379	0.1418	0.1405	0.1419	0.1377	0.1033	0.0879	0.0756	0.1065	0.1078	0.1107	0.1106	0.1135	0.1028	0.0991	0.1021	0.1016	0.1001	0	0.0026	0.0065	0.0036	0.0027	0.0046	0.0061
ASW	0.1254	0.1224	0.1263	0.1248	0.1262	0.119	0.0829	0.0677	0.0566	0.0861	0.0875	0.0899	0.0899	0.0928	0.0841	0.0802	0.0839	0.0835	0.0819	0.0026	0	0.0097	0.0098	0.0089	0.0103	0.0107
LWK	0.1524	0.1499	0.1538	0.1525	0.1538	0.1512	0.1178	0.103	0.0905	0.1232	0.1244	0.1278	0.1276	0.1304	0.1169	0.1135	0.1158	0.1152	0.1138	0.0065	0.0097	0	0.008	0.0073	0.0096	0.0109
ESN	0.1631	0.1608	0.1647	0.1634	0.1648	0.1619	0.1289	0.114	0.101	0.1354	0.1367	0.14	0.1398	0.1425	0.1287	0.1251	0.1274	0.1268	0.1252	0.0036	0.0098	0.008	0	0.0009	0.0053	0.0075
YRI	0.1613	0.1589	0.1627	0.1614	0.1628	0.1599	0.1272	0.1125	0.0995	0.1336	0.1349	0.1383	0.1381	0.1407	0.1271	0.1237	0.1259	0.1252	0.1237	0.0027	0.0089	0.0073	0.0009	0	0.004	0.0062
MSL	0.1636	0.1611	0.1651	0.1638	0.1652	0.162	0.1282	0.1135	0.1004	0.1352	0.1365	0.1398	0.1396	0.1423	0.1286	0.1249	0.1272	0.1265	0.1249	0.0046	0.0103	0.0096	0.0053	0.004	0	0.0037
GWD	0.1581	0.1557	0.1595	0.1582	0.1595	0.1567	0.1238	0.1091	0.0962	0.1295	0.1308	0.1342	0.1341	0.1368	0.1236	0.1201	0.1224	0.1217	0.1204	0.0061	0.0107	0.0109	0.0075	0.0062	0.0037	0

Supplementary Information Table 5: F_{ST} between population pairs. Pairwise F_{ST} between populations was calculated using Weir and Cockerham's estimator. F_{ST} was estimated for each chromosome separately, and then averaged across chromosomes. The highest FST value within each continental group is highlighted in red.

Imputation Reference Panel	Function	Genotyped and/or Imputed				Imputed			
		Total	Rare Freq. < 0.5%	Low-Frequency 0.5% <= Freq. < 5%	Common Freq. >= 5%	Total	Rare Freq. < 0.5%	Low-Frequency 0.5% <= Freq. < 5%	Common Freq. >= 5%
Genotyped (after QC)	Protein-Altering SNV	5,966	0	322	5,644	…	…	…	…
	Other SNV	309,862	0	8,247	301,615	…	…	…	…
	TOTAL	**315,905**	**0**	**8,570**	**307,335**	…	…	…	…
HapMap Phase 2 Release 22 : CEU	Protein-Altering SNV	13,375	190	2,347	10,838	7,332	190	2,024	5,118
	Other SNV	2,346,036	20,180	284,594	2,041,262	2,036,174	20,180	276,347	1,739,647
	TOTAL	**2,359,411**	**20,370**	**286,941**	**2,052,100**	**2,043,506**	**20,370**	**278,371**	**1,744,765**
HapMap Phase 3 Release 2 : CEU+TSI	Protein-Altering SNV	11,091	654	2,141	8,296	5,048	654	1,818	2,576
	Other SNV	1,204,666	24,795	115,132	1,064,739	894,804	24,795	106,885	763,124
	TOTAL	**1,215,757**	**25,449**	**117,273**	**1,073,035**	**899,852**	**25,449**	**108,703**	**765,700**
1000 Genomes Project Phase 1 Release 3	Protein-Altering SNV	57,430	22,509	16,509	18,412	51,387	22,509	16,186	12,692
	Other SNV	13,565,333	4,107,239	3,370,388	6,087,706	13,255,471	4,107,239	3,362,141	5,786,091
	Protein-Altering Indels	1,143	218	355	570	1,143	218	355	570
	Other Indels	502,558	91,068	124,660	286,830	502,558	91,068	124,660	286,830
	Structural Variation	372	194	82	96	372	194	82	96
	TOTAL	**14,126,836**	**4,221,228**	**3,511,994**	**6,393,614**	**13,810,931**	**4,221,228**	**3,503,424**	**6,086,279**
1000 Genomes Project Phase 3 Release 5	Protein-Altering SNV	68,363	33,099	16,569	18,695	62,397	33,099	16,247	13,051
	Other SNV	16,068,577	6,079,649	3,586,656	6,402,272	15,758,715	6,079,649	3,578,409	6,100,657
	Protein-Altering Indels	2,157	896	544	717	2,157	896	544	717
	Other Indels	851,262	234,344	168,010	448,908	851,262	234,344	168,010	448,908
	Structural Variation	5,836	2,603	1,418	1,815	5,836	2,603	1,418	1,815
	TOTAL	**16,996,195**	**6,350,591**	**3,773,197**	**6,872,407**	**16,680,367**	**6,350,591**	**3,764,628**	**6,565,148**

Supplementary Information Table 6: Summary of imputed variants in AMD GWAS. The count of imputed variants includes only those variants that were polymorphic after imputation and for which estimated imputation quality passed standard thresholds ($r^2 > 0.30$). These filters focus attention on European ancestry variants and account for most of the difference between the number of variants characterized in the 1000 Genomes panel and those available for analysis after imputation.

Imputation Reference Panel	Number of Genotyped and Well-imputed (R^2 >= 0.3) Variants	Observations from 1,000 Permutations		
		Range of Top P Values	50th Smallest Top P Value	Effective Number of Independent Tests (Bonferroni correction)
HapMap Phase2 Release 22 (CEU; N =60)	2,359,411	[5.3×10^{-10} ; 9.1×10^{-6}]	6.23×10^{-8}	802,251
1000 Genomes Phase 1 Release 3 (ALL; N = 1,092)	14,120,996	[4.3×10^{-11} ; 3.0×10^{-6}]	1.31×10^{-8}	3,820,025
1000 Genomes Phase 3 Release 5 (ALL; N = 2,504)	16,806,192	[8.3×10^{-12} ; 1.9×10^{-6}]	1.47×10^{-8}	3,410,725

Supplementary Information Table 7: GWAS type 1 error simulation. For each imputed dataset we performed 1,000 genome-wide association analyses (score test) after reshuffling the case-control status of 2,136 AMD cases and 1,139 controls. We sorted the top p value of each of the 1000 analyses and empirically determined the significance threshold (alpha = 5%) as the P value of the 50th smallest observed GWAS top hit.

Chr.	Position	dbSNP ID	Gene	Amino Acid Change	Minor Allele Frequency*	Imputation Reference Panel					
						HapMap2 Release 22		1000 Genomes Project Phase 1 Release 3		1000 Genomes Project Phase 3 Release 5	
						Rsq	*P*	Rsq	*P*	Rsq	*P*
1	196,659,237	rs1061170	*CFH*	p.Y402H	38.20%	0.76	1.16×10^{-66}	0.997	1.35×10^{-76}	0.997	1.97×10^{-76}
6	31,914,024	rs4151667	*CFB*	p.L9H	4.20%	...	...	0.993	3.80×10^{-10}	0.994	3.94×10^{-10}
6	31,914,180	rs641153	*CFB*	p.R32Q	8.70%	...	...	0.965	2.86×10^{-18}	0.984	4.27×10^{-19}
10	124,214,448	rs10490924	*ARMS2*	p.A69S	20.60%	0.797	1.44×10^{-61}	0.887	3.85×10^{-68}	0.884	2.90×10^{-67}
19	6,718,387	rs2230199	*C3*	p.R102G	20.90%	0.664	1.62×10^{-10}	0.738	3.74×10^{-12}	0.718	1.61×10^{-12}
19	45,411,941	rs429358	*APOE*	p.C130R	11.70%	...	...	0.857	1.55×10^{-06}	0.824	1.39×10^{-06}

Supplementary Information Table 8: Summary of Genome-wide Association Results for AMD GWAS, summarizing imputation quality (r^2) and association p-values for putatively functional variants. *Minor allele frequency from NHLBI ESP Exome Variant Server (EVS, release ESP6500).

Locus	Reference Panel	Chr.	95% Credible Set							Top Variant			
			Interval			Number of Variants							
			Start Position	End Position	Size [bp]	SNPs	Indels	Structural Variation	Total	Position	Annotation	*P*	Posterior Probability
#1	HapMap / Phase 2	1	196,679,455	196,702,810	23,356	6	0	0	6	196,679,455	*CFH* / Intron	5.0×10^{-19}	36.80%
	1000 Genomes / Phase 1	1	196,679,455	196,704,632	25,178	19	0	0	19	196,704,632	*CFH* / Intron	7.6×10^{-21}	17.00%
	1000 Genomes / Phase 3	1	196,679,455	196,704,632	25,178	19	1	0	20	196,704,632	*CFH* / Intron	2.2×10^{-19}	15.60%
#2	HapMap / Phase 2	1	196,646,176	196,672,473	26,298	14	0	0	14	196,646,176	*CFH* / Intron	5.6×10^{-18}	26.20%
	1000 Genomes / Phase 1	1	196,646,176	196,704,997	58,822	73	8	0	81	196,664,082	*CFH* / Intron	7.4×10^{-18}	7.50%
	1000 Genomes / Phase 3	1	196,646,176	196,704,997	58,822	76	8	0	84	196,646,261	*CFH* / Intron	1.5×10^{-17}	3.90%
#3	HapMap / Phase 2	6	31,930,462	31,930,462	1	1	0	0	1	31,930,462	*SKIV2L* / Intron	1.1×10^{-21}	99.80%
	1000 Genomes / Phase 1	6	31,894,355	31,930,462	36,108	2	0	0	2	31,930,462	*SKIV2L* / Intron	1.0×10^{-21}	65.50%
	1000 Genomes / Phase 3	6	31,894,355	31,930,462	36,108	2	0	0	2	31,930,462	*SKIV2L* / Intron	1.1×10^{-21}	79.10%
#4	HapMap / Phase 2	10	124,215,315	124,219,275	3,961	3	0	0	3	124,219,275	Intergenic	9.2×10^{-69}	85.20%
	1000 Genomes / Phase 1	10	124,210,369	124,226,630	16,262	16	1	1*	18	124,215,565	*ARMS2* / Intron	8.1×10^{-74}	8.00%
	1000 Genomes / Phase 3	10	124,211,536	124,226,630	15,095	20	1	1*	22	124,216,824	*ARMS2 / SV*	3.7×10^{-73}	10.20%
#5	HapMap / Phase 2	19	6,718,387	6,724,340	5,954	2	0	0	2	6,718,387	C3 : p.(Arg102 Gly)	9.1×10^{-11}	74.00%
	1000 Genomes / Phase 1	19	6,713,175	6,718,387	5,213	3	1	0	4	6,718,387	C3 : p.(Arg102 Gly)	1.7×10^{-12}	61.80%
	1000 Genomes / Phase 3	19	6,713,175	6,722,817	9,643	3	2	0	5	6,718,387	C3 : p.(Arg102 Gly)	6.6×10^{-13}	55.40%

Supplementary Information Table 9: Bayesian Credible Sets for AMD GWAS Including Putatively Functional Variants. * NM_001099667.1:c.(*)372_815del443ins54

Continental Grouping	No Proxies	Average Number of Proxies	Percentiles			
			5%	Q1	Q3	95%
AFR	22.5%	14.4	0	1	11	55
AMR	10.4%	30.3	0	3	29	101
EAS	11.2%	44.4	0	4	42	170
EUR	8.8%	38.2	0	4	37	128
SAS	10.9%	31.8	0	3	32	105
All of 1000 Genomes	16.6%	16.6	0	1	17	63
Intersection	34.9%	8.2	0	0	7	31
Union	2.9%	66.3	2	2	66	231

Supplementary Information Table 10: Number of Proxies for GWAS Catalog Loci. This analysis uses a thinned set of 3,990 loci reaching $p < 5x10^{-8}$ in the GWAS catalog. For thinning, only one variant was considered whenever two or more catalog variants were in $r^2 > 0.80$ with each other in the combined 1000 Genomes data set. For each variant in the thinned set we calculated the number of proxies (variants in $r^2 > 0.80$) in each continental grouping and in the full 1000 Genomes Project dataset. Then, we calculated the union list of proxies (variants that were considered proxies in at least one continental groupings) and the intersection list of proxies (variants that were considered proxies in all continental groupings). The table summarizes key statistics for this distribution, including the proportion of variants with no proxies at all, the average number of proxies per variant, and percentiles of the proxy count distribution (5^{th}, Q1 / 25^{th}, Q3 / 75^{th} and 95^{th}).

							Sequenced samples with OMNI2.5 genotypes			Sequenced samples with AFFY6.0 genotypes (not in OMNI2.5)		
Continental Group	**Population**	**# sequenced samples**	**# samples with OMNI2.5**	**# samples with AFFY6.0**	**# sequenced samples with OMNI2.5**	**# sequenced samples with AFFY6.0 but not OMNI2.5**	**Trio phased**	**Duo phased**	**Unrelated**	**Trio phased**	**Duo phased**	**Unrelated**
AMR	ACB	96	102	124	77	19	45	0	32	4	0	15
	CLM	94	107	152	67	27	65	1	1	11	0	16
	MXL	67	103	5	66	1	60	2	4	0	0	1
	PEL	86	105	130	69	17	69	0	0	6	0	11
	PUR	105	111	156	71	34	71	0	0	15	0	19
	ASW	66	104	110	60	6	26	20	14	0	0	6
SAS	BEB	86	0	157	0	86	0	0	0	45	8	33
	GIH	106	113	62	106	0	0	5	101	0	0	0
	ITU	103	0	139	0	103	0	0	0	4	4	95
	PJL	96	0	162	0	96	0	0	0	76	7	13
	STU	103	0	176	0	103	0	0	0	14	5	84
EAS	CDX	99	100	49	97	2	0	3	94	0	0	2
	CHB	103	108	5	103	0	0	0	103	0	0	0
	CHD	0	1	0	0	0	0	0	0	0	0	0
	CHS	108	153	157	96	12	96	0	0	12	0	0
	JPT	104	105	5	104	0	0	0	104	0	0	0
	KHV	101	121	3	101	0	41	1	59	0	0	0
EUR	CEU	99	183	73	99	0	90	5	4	0	0	0
	IBS	107	150	30	100	7	100	0	0	7	0	0
	FIN	99	100	37	96	3	0	0	96	0	0	3
	GBR	92	104	26	92	0	0	2	90	0	0	0
	TSI	108	112	12	108	0	0	0	108	0	0	0
AFR	ESN	99	0	238	0	99	0	0	0	79	16	4
	GWD	113	0	301	0	113	0	0	0	113	0	0
	LWK	101	116	11	101	0	0	3	98	0	0	0
	MKK	0	31	0	0	0	0	0	0	0	0	0
	MSL	85	0	141	0	85	0	0	0	51	15	19
	YRI	109	189	25	109	0	99	2	8	0	0	0
	Total	**2535**	**2318**	**2486**	**1722**	**813**	**762**	**44**	**916**	**437**	**55**	**321**

Supplementary Information Table 11: Input data to haplotype scaffold construction.

Individual	Haplotype Concordance	Switch Error Rate	Flip Error Rate	Mean inter-switch distance (kb)	Mean length of incorrectly phased haplotype (kb)
NA19240	99.40%	0.34%	0.31%	2406.6	27.4
HG02799	98.44%	0.69%	0.64%	1010.1	29.6
HG03108	99.01%	0.50%	0.47%	1875.1	21.6
NA12878	98.59%	0.74%	0.62%	846.6	19.8
HG03428	90.17%	0.55%	0.43%	149.3	70.0
NA20847	92.82%	0.56%	0.39%	85.1	54.8
Average	**96.41%**	**0.56%**	**0.48%**	**1,062.1**	**37.2**

Supplementary Information Table 12: Comparison to fosmid phasing**.** Switch error statistics between fosmid and 1000 Genomes haplotypes. Flip errors refer to individual alleles appearing on the opposite haplotype, and do not contribute to inter-switch distances.