The Canadian Longitudinal Study on Aging

Genome-wide Genetic Data Release (version 3)

Vincenzo Forgetta[1], Corinne Darmond-Zwaig[2], Alexandre Belisle[2], Rui Li[2], Cynthia Balion[3], Delnaz Roshandel[4], Christina Wolfson[5], Susan Kirkland[6], Guillaume Lettre[7], Guillaume Pare[3], Andrew D. Paterson[4], Chris Verschoor[8], Mark Lathrop[2], Parminder Raina[8], J. Brent Richards[1,9,10], and Jiannis Ragoussis[2,11]

[1]Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal, QC, Canada, [2]McGill University and Genome Quebec Innovation Centre, Department of Human Genetics, McGill University, Montreal, QC, Canada, [3]Hamilton Regional Laboratory Medicine Program, McMaster University, St. Joseph's Hospital St. Luke's Wing, Hamilton, ON, Canada, [4]The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, ON, Canada, [5]Department of Epidemiology and Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada, [6]Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada, [7]Montreal Heart Institute and Universite de Montreal, Montreal, QC, Canada, [8]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada, [9]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK, [10]Department of Human Genetics, McGill University, Montreal, QC, Canada, [11]Department of Bioengineering, McGill University, Montreal, QC, Canada

Abstract

Genetic and environmental factors contribute to maintaining health and in the development of disease and disability as people age. The Canadian Longitudinal Study on Aging (CLSA) is a national long-term cohort study that will follow approximately 50,000 men and women, and presents a unique opportunity to study genetic and environmental contributions to human health and disease by providing information on the changing biological, medical, psychological, social, lifestyle and economic aspects of participants' lives. This document describes the availability and quality assessment of genetic data for 26,622 CLSA participants, comprising genome-wide genotype data for 794,409 markers and whole-genome imputed data for ~308 million genetic variants. Quality assessment includes both marker- and sample-based tests, as well as analysis of population structure and familial relatedness. Qualified researchers from any country can obtain access to these genomic and phenotypic data via the CLSA data access webpage at https://www.clsa-elcv.ca/data-access.

Data Release Summary

- Genotype data for 26,622 CLSA participants with 93% of European ancestry.
- Affymetrix Axiom array genotypes for 794,409 genetic variants, of which 95% are high quality.
- TOPmed imputed genotypes for ~308 million genetic variants.

Contents

1 Introduction

1.1 The CLSA cohort

The Canadian Longitudinal Study on Aging (CLSA) is a large, national, long-term cohort study that will follow approximately 50,000 men and women who are between the ages of 45 and 85 years when recruited, for at least 20 years. The CLSA will collect information on the changing biological, medical, psychological, social, lifestyle and economic aspects of people's lives. These factors will be studied to understand how, individually and in combination, they have an impact in both maintaining health and in the development of disease and disability as people age. Recruitment and baseline data collection took place 2010-2015. The first follow up was completed 2015-2018.
In addition to this, DNA extraction and genotyping was performed at the McGill and Genome Quebec Innovation Centre, Montreal, Canada. Participants were genotyped using the Affymetrix UK Biobank Axiom array [1].

1.2 The UK Biobank Axiom genotyping array

The UK Biobank Axiom array was used to genotype ~450,000 individuals in the UK Biobank cohort [2]. The array has been designed to target known disease associated variants, known coding variants, as well as a panel of variants for optimal imputation in individuals of European ancestry, an ancestry that includes over 90% of CLSA genotyped participants (see section 3.1). For more information regarding the UK Biobank Axiom genotyping array, please refer to [1].

1.3 About this data release

This data release contains genotype data for 26,622 successfully genotyped CLSA participants across 794,409 genetic markers, as well as 308 million genetic variants imputed from the TOPMed reference panel [3]. The marker- and sample-based quality control presented in this document largely follows the same procedure used by UK Biobank [4]. Genomic positions of the array genotyped and imputed genotype data are reported in reference to human genome build GRCh37/hg19 and GRCh38/hg38, respectively. In addition, this data release contains genetic data for the control samples used during array genotying (section 1.4.2 and section 1.4.3).

1.3.1 CLSA Genetic Data Files and File Formats

The following files are provided with this data release:

- Filename: `clsa_gen_v3.bed` — PLINK binary biallelic genotype table.
- Filename: `clsa_gen_v3.bim` — PLINK extended MAP file.
- Filename: `clsa_gen_v3.fam` — PLINK sample information file. The individual identifier (IID) corresponds to the `ADM_GWAS_COM` variable, which is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments. Also included are control samples used during array genotyping (section 1.4.2 and section 1.4.3).

The directly genotyped data are in binary PLINK format (`.bed`, `.bim`, `.fam`). For more information, please refer to https://www.cog-genomics.org/plink/1.9/formats. The `.bim` and `.fam` file lists the order of the markers and genotyped individuals, respectively. It is recommended to use PLINK to manipulate these files (https://www.cog-genomics.org/plink/1.9/).

- Filename: `clsa_imp_{1..23}_v3.bgen` — Genotype imputation for the autosomes and X chromosome (coded as 23).
- Filename: `clsa_imp_{1..23}_v3.bgen.bgi` — `.bgen` index file for the autosomes and X chromosome (coded as 23).
- Filename: `clsa_imp_v3.sample` — Sample information file for imputed genotype data. The individual identifier (`ID_1`) corresponds to the ADM_GWAS_COM variable, which is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments. Also included are control samples used during array genotyping (section 1.4.2 and section 1.4.3).
- Filename: `clsa_mfi_{1..23}_v3.txt` — Contains alternate allele frequency and imputation quality score (`Rsq`) for each of the markers in the imputed data. Columns are: `rsid`, `ref_allele`, `alt_allele`, `alt_allele_frequency.1`, `Rsq.1`, `alt_allele_frequency.2`, `Rsq.2`. Two values for alternate allele frequency and imputation quality score are provided, one for each batch of the imputation process (see section 5).

The imputed genotyped data are in binary BGEN version 1.2 format using 8-bit encoding [5]. The `.sample` file lists the order of the individuals within the imputed genotype data files (`.bgen`). It is recommended to use `qctool` version 2 [6] or `bgenix` [7] to manipulate this data type. A list of other compatible software is listed on the BGEN format website [8].
The conversion of the original VCF files (1.9 TB) to the more efficient BGEN format (187 GB) using PLINK2 resulted in collapsing of the raw genotype probabilities down to dosage. Therefore, lossless extraction of the original genotype probabilities is not possible using the BGEN files provided [9]. This is of limited consequence as most downstream analysis use dosage only, such as most modern GWAS software. Should your analyses require the original genotype probabilities, requests for the original VCF files can be directed to the contact information provided on the CLSA data access webpage at https://www.clsa-elcv.ca/data-access.

- Filename: `clsa_mqc_v3.txt` — Marker QC metrics file

The marker quality control metrics file contains information pertaining to the quality assessment of the directly genotyped markers (section 3). It is a plaintext file with the following space separated columns:

1. `affymetrix_probeset_id` — datatype: string — Identifier for probeset from Affymetrix array definition.
2. `affymetrix_snp_id` — datatype: string — Identifier for marker from Affymetrix array definition.
3. `rs_id` — datatype: string — If available, the dbSNP identifier.
4. `chromosome` — datatype: numeric — Chromosome label 1-26; 23=X,24=Y,25=XY,26=MT.
5. `position` — datatype: numeric — Physical position of start of marker/Indel in GRCh37 coordinates.
6. `allele_1` — datatype: string — Marker allele 1.
7. `allele_2` — datatype: string — Marker allele 2.
8. `batch_disc` — datatype: numeric — Count of batches where marker failed genotype frequency discordance test.
9. `hwe_disc` — datatype: numeric — Count of batches where marker failed HWE test.
10. `ctl_disc` — datatype: numeric — Count of batches where marker failed control genotype discordance test.
11. `sex_disc` — datatype: numeric — Count of batches where marker failed sex genotype frequency discordance test.

12. `low_freq` — datatype: (0/1) [no/yes] — Marker has low minor allele frequency.
13. `indel` — datatype: (0/1) [no/yes] — Marker is an insertion or deletion.

- Filename: `clsa_sqc_v3.txt` — Sample QC metrics file.

The sample quality control metrics file contains information pertaining to the quality assessment per genotyped sample (section 4). It is a plaintext file with the following space separated columns:

1. `ADM_GWAS_COM` — datatype: numeric — This identifier is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments.
2. `batch` — datatype: numeric — Indicates the batch from which the called the genotypes for this sample originated.
3. `selfreported.sex` — datatype: (0/1/2) [missing/male/female] — Self-reported sex.
4. `chromosomal.sex` — datatype: (0/1/2) [missing/male/female] — Sex chromosome composition as determined via PLINK (section 1.4.4).
5. `pca.cluster.id` — datatype: numeric — Indicates cluster number based on a principal components analysis of the genotypes (section 4.5).
6. `in.kinship` — datatype: (0/1) [no/yes] — Indicates samples which have at least one relative of 3rd degree or closer among the set of genotyped individuals (section 4.2).
7. `in.hetmiss` — datatype: (0/1) [no/yes] — Indicates samples identified as outliers in heterozygosity or genotype missingness (section 4.3).

8-27. `PC1-20` — datatype: numeric — Score for each principal component 1-20 (section 4.4).

28-47. `ePC1-20` — datatype: numeric — Score for each principal component 1-10 for analysis of individuals within the European ancestry subset only (section 4.5).

- Filename: `clsa_rel_v3.txt` — Sample relatedness file.

The sample relatedness file lists the pairs of individuals related up to the third degree in the data set. The file contains information obtained from the analysis of familial relatedness (see section 4.2) using the KING software program [10].

1. `ADM_GWAS_COM_1` — datatype: numeric — This identifier is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments.
2. `ADM_GWAS_COM_2` — datatype: numeric — This identifier is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments.
3. `HetHet` — datatype: numeric — Proportion of SNPs with double heterozygotes (e.g., AG and AG).
4. `IBS0` — datatype: numeric — Porportion of SNPs with zero IBS (identical-by-state) (e.g., AA and GG).
5. `Kinship` — datatype: numeric — Estimated kinship coefficient (ϕ) from the SNP data.
6. `InfType` — datatype: string — Inferred relationship type, such as Dup/MZTwin, PO, FS, 2nd, 3rd, 4th, UN.

- Filename: `clsa_hla_v3.csv` — The imputation of HLA alleles.

The HLA imputation file contains information pertaining to the imputation of classical human leukocyte antigen (HLA) alleles from SNP genotypes (**??**). It is a plaintext file with the following space separated columns:

1. ADM_GWAS_COM — datatype: numeric — This identifier is used as a link to project-specific CLSA participant identifiers associated with data from questionnaires and physical assessments.
2. Locus — datatype: string — The HLA gene type identifier.
3. Allele1 — datatype: numeric — one allele group in 4-digit format.
4. Allele1_Probability — datatype: numeric — The individual probability that the call of allele1 is correct.
5. Allele2 — datatype: numeric — The other allele group in 4-digit format.
6. Allele2_Probability — datatype: numeric — The individual probability that the call of allele2 is correct.
7. Combined_Probability — datatype: numeric — The overall probability that both calls have been assigned correctly.

1.4 DNA extraction, genome-wide genotyping, and quality filtering

1.4.1 Sample storage and DNA extraction

Whole blood buffy coats were isolated from peripheral blood drawn into EDTA vacutainers following centrifugation at 2000 $\times$ g for 10 minutes and removal of the plasma layer. Buffy coats from three 6 ml EDTA vacutainers were pooled and 200μl aliquots transferred to 2D barcoded screw-top storage tubes. Samples were immediately moved to -80°C storage, and transferred to liquid N_2 storage for up to one week later until shipment to the genomics facility, after which they were stored at -20°C. The time from blood collection to -80°C storage was under two hours for all participants. Genomic DNA was extracted from blood samples using the purification protocol "Chemagic DNA Buffy Coat Kit special 200μl prefilling VD151007" on the Chemagic MSM I instrument (Perkin-Elmer article No. CMG-533). The reagents needed for the extraction were included in the Chemagic Buffy Coat Kit (Perkin-Elmer article No. CMG-713). Samples were eluted into 200μl/well of 10mM Tris-HCl pH 8.0. All extracted samples were quantified using PicoGreen Reagent Kit (Life Technologies, catalog # P7589). A minimum concentration for passing of samples was set at 10 ng/μl. Samples were subsequently normalized to 20 ng/μl, except for those with a concentration of 10-20 ng/μl, which were used undiluted. Each sample was aliquoted into 3 separate Samplosophy®2D barcoded tubes from LVL technologies. Two of the aliquots were stored at -80°C for long term storage. The last aliquot was stored at -20°C and used to create normalized working plates.

1.4.2 Genotyping and calling

Each plate genotyped contained 92 CLSA DNA samples and 4 controls. Position A01 on each plate contained the Affymetrix Reference Genomic DNA 103 (Catalog# 900421) or Personal Genome Project NA24385 (male). Positions D06 and H10 contained the CEPH control 1463-02 (Coriell Cell Repositories, catalog # NA12878, female Caucasian) or 1347-2 (Coriell Cell Repositories, catalog # NA10859, female). Position G11 on each plate was reserved for a deionized water negative control. For sample preparation, including whole genome amplification, fragmentation, precipitation and re-suspension, and hybridization to the array, the Affymetrix protocol (Axiom 2.0 Assay Automated workflow on Affymetrix NIMBUS) was followed. Samples were hybridized to UK Biobank arrays (Thermo Fisher Catalog # 902502). Axiom Array plates were processed on the Affymetrix GeneTitan Multi-Channel (MC) Instrument. For first pass QC, batches of 8 plates were analyzed using the Sample QC workflow of the Axiom™ Analysis Suite 2.0 software where a subset of 20,000 reliable probes are used to determine Sample QC and Dish QC, the later being the measure of the resolution of the AT and GC signal contrast.

1.4.3 Genotyping quality control and removal of duplicate genotyped participants

Genotyping was undertaken in 5 separate batches of roughly 5,000 samples each using Axiom™ Analysis Suite 2.0, similar to UKBiobank genotyping QC documentation [11]. Genotype calling resulted in 27,010 successfully genotyped DNA samples. An inclusion SNP list containing 794,409 genetic variants was used [1], as well as the following QC parameters for selecting samples passing to further analysis as follows (default parameters are in brackets):

- Dish QC $\geq$ 0.82 (0.82)
- QC call rate $\geq$ 95.0 (97.0)
- Percent of passing samples $\geq$ 70.0 (95.0)
- Average call rate for passing samples $\geq$ 95.0 (98.5)

During the genotyping process, DNA plates or samples with genotype calling quality lower than the thresholds set above were re-genotyped, resulting in some individuals being genotyped twice. Duplicate genotyped participants were detected by KING version 2.1.3 [10] using the following sequential steps:

- Removed SNPs with minor allele frequency (MAF) $<$ 0.05 in both batches.
- Merged batches into a single dataset using PLINK.
- Used the KING software program to detect duplicates (`--duplicate` parameter).
- For a given control sample, ensured that all duplicates detected were the same controls on another DNA plate.
- For duplicate CLSA samples:
  - Ensured that the participant identifiers were identical.
  - Removed the sample with higher genotype missingness.

Using the above procedure to remove duplicates resulted in 26,622 uniquely genotyped CLSA participants. All subsequent analyses, outlined below, were performed on this de-duplicated set of 26,622 genotyped CLSA participants. For each of the three controls (see section 1.4.2), the genotyped sample with the highest calling rate was retained as part of this genetic data release.

1.4.4 Sex chromosome composition

Sex defined by sex chromosome number was determined using PLINK version 1.9 [12] [13] using the following sequential steps:

- Removed SNPs with MAF $<$ 0.05.
- Retained linkage disequilibrium ($r^2 < 0.5$) pruned SNPs on chromosome X and Y only using window size of 20,000 markers and step size 2,000 markers (PLINK option `--indep-pairphase 20000 2000 0.5`).
- Determine sex chromosome composition by estimating heterozygisoty of X chromosome markers and counting Y chromosome markers using PLINK (`--check-sex ycount 0.4 0.8`).

Distribution of chromosome X F estimates showed a gap between 0.4 and 0.8 (Figure 1). Using this threshold, we obtained sex chromosome number and compared this to self-reported gender (Table 1). Using sex chromosome composition we then set all heterozygous haploid and non-male Y chromosome genotype calls to missing (PLINK `--set-hh-missing` option). All subsequent analyses in this document will use sex chromosome composition where required.

Self-reported Gender	Sex Chromosome Composition	Count
Male	Male	13324
Female	Female	13250
Female	Male	17
Male	Female	16
Female	Missing	10
Male	Missing	5

Table 1: Count of individuals by self-reported gender and sex chromosome composition.

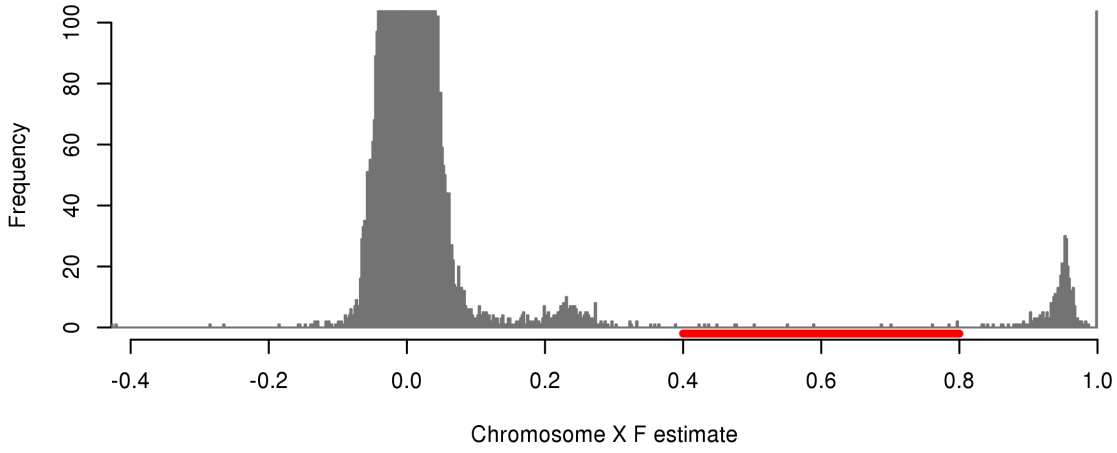

Figure 1: Distribution of chromosome X F estimates for CLSA genotyped participants (y-axis truncated). Individuals with chromosome X F estimates within the range of 0.4 to 0.8 (red) are considered to have undefined chromosomal sex.

2 HLA Type Imputation

We used the HLA*IMP:02 method [14] and a multi-population reference panel [14] (ThermoFisher Catalog # 000.911) to impute HLA types. The genotypes of 11 major MHC Class I and Class II loci with 4-digit resolution were imputed for HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4, -DRB5. For the positive controls, the imputation was done for 587 replicates of NA12878, 75 replicates of NA24385 and 4 replicates of NA10859. The alleles called with a posterior probability threshold as 0.7 were compared to their known genotypes from literature [15, 16, 17, 18, 19]. Calling accuracy was 100% across the loci. The imputation accuracy of genotyped CLSA participants was estimated by using the replicated samples. The validation rate was 100% for all the replicates.

3 Marker-based quality control

The quality control analysis consisted of 4 tests intended to check for consistency of markers across various experimental factors, such as genotyping batch, participant sex, Hardy-Weinberg equilibrium (HWE), and discordance of genotyping across control replicates. All tests were conducted on a subset of ancestrally homogeneous participants (see section 3.1). Details of each test are described in section 3.2. The purpose of the marker-based quality control is not to exclude SNPs for future analyses. Rather, it is to flag SNPs which may have low quality. Therefore all SNPs are available for researchers, but the analysis below identifies SNPs failing the tested QC parameters.

3.1 Subset for ancestrally homogeneous participants

The QC tests are only effective in the context of single ancestral population, such as individuals of European ancestry. To address this requirement, the marker QC tests were performed on the largest subset of ancestrally homogeneous participants. The following describes the procedure used to identify these participants.
The largest ancestral group of participants was determined via comparison to 414 individuals across 4 populations from 1000 Genomes phase 3. These 4 ancestral groups represent individuals of Northern and Western European (CEU), Han Chinese (CHB), Japanese (JPT) and Yoruban (YRI) ancestry. We first extracted the Affymetrix UK Biobank Axiom array markers (N=794,409) from 1000 Genomes [20] and retained markers with the following criteria:

- MAF > 0.05.
- Hardy-Weinberg equilibrium (HWE) p-value $> 10^{-6}$.
- Single nucleotide substitutions with single character allele-codes (A, C, G, or T) (PLINK option `--snps-only just-acgt` option).
- Markers with unambiguous strand, that is exclude A/T or C/G markers.
- Retained linkage disequilibrium ($r^2 < 0.1$) pruned markers using window size of 1000 kb and step size of 5 markers (PLINK option `--indep-pairwise 1000 kb 5 0.1`).

This resulted in a set of 45,251 markers from which we computed principal component loadings on the 414 individuals from 1000 Genomes phase 3. We then projected the CLSA samples onto the principal components of the 1000 Genomes analysis. K-means clustering of the top 4 PCs was used to select the largest cluster, which was found to overlap the cluster containing CEU population (Figure 2). There were 4,500, 4,570, 4,351, 4,471, and 6,469 samples in largest clusters in batch 1 to 5, respectively. The sum of CLSA participants in the largest clusters from all batches represented 91% of the total number of individuals genotyped.

3.2 Marker-based quality control tests

The marker quality control tests were performed on a subset of 24,358 ancestrally homogeneous participants (see section 3.1) typed at 794,409 markers. We conducted 4 tests to check for marker consistency across various experimental factors:

1. Discordance of genotype frequency between batches.
2. Departure from Hardy-Weinberg equilibrium (HWE).

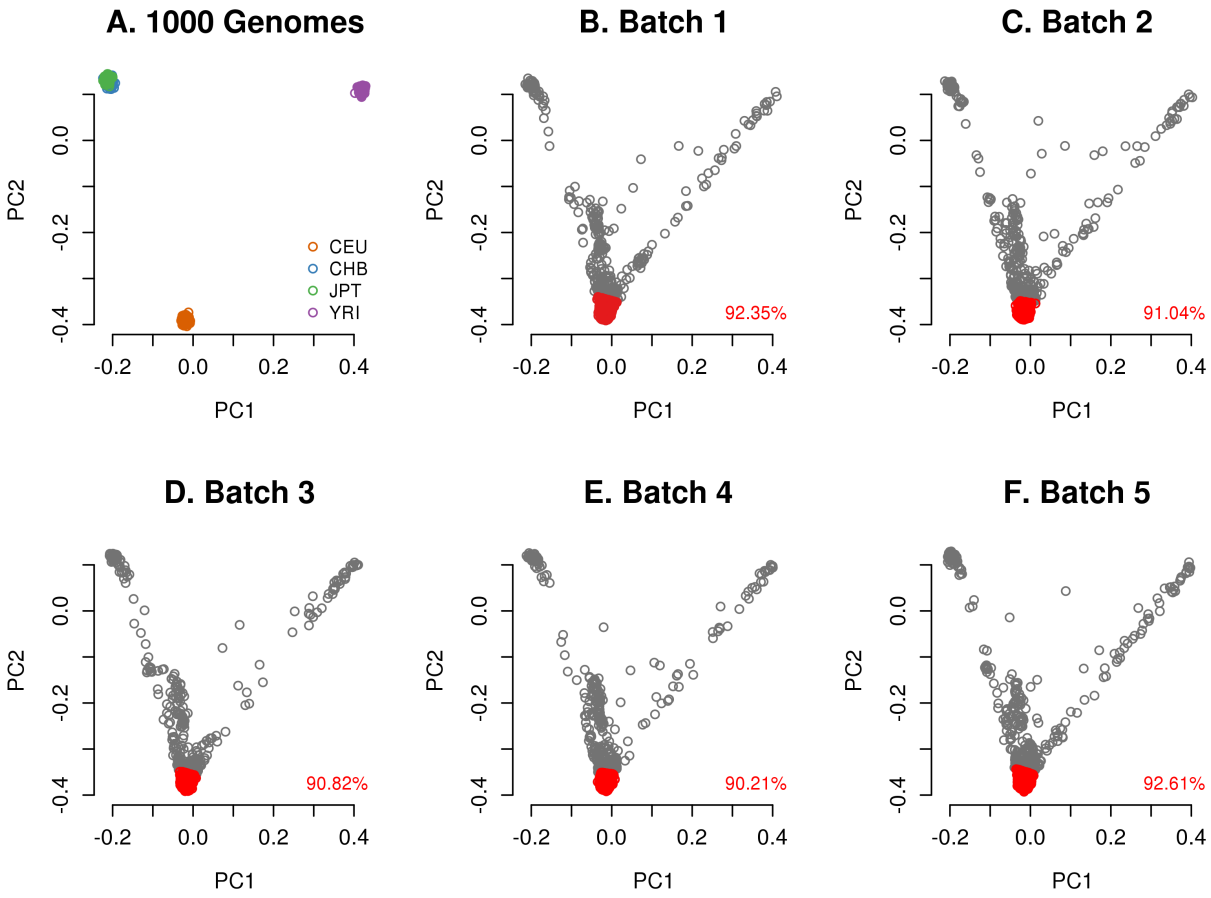

Figure 2: Principal component (PC) plots. (A) Plot of first 2 PC for the analyzed populations from 1000 Genomes. (B-F) Projection of CLSA participants onto 1000 Genomes PC plot for genotype batch 1 to 5 followed by k-means clustering of PC1-4 (grey points). The largest cluster overlaps the 1000 Genomes CEU population (red points and percentage of total in batch is provided).

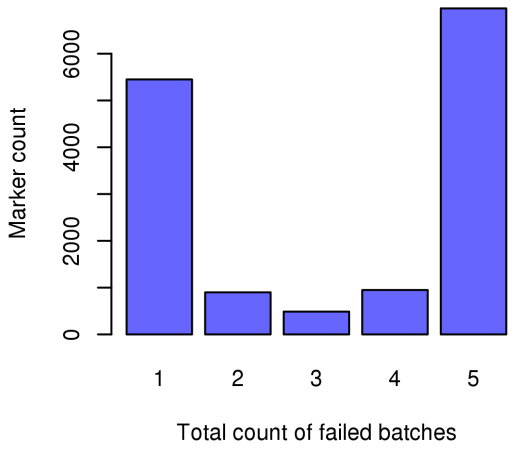

Figure 3: Count of markers that failed discordant genotype frequency test in one or more batches.

3. Discordance across control replicates.
4. Discordance in genotype frequency between males and females.

The status of each marker in each test is noted in the marker quality control information file.

3.2.1 Choosing a p-value threshold for the marker quality control tests

We determined a multiple-testing corrected p-value threshold for quality control tests as 3.15×10^{-10}. For the 794,409 markers and 5 batches, this p-value cut-off can be considered as a family-wise error rate of 0.0013 for each test. Since many tests may be positively correlated, the threshold is conservative and will identify markers with strong evidence of deviation from the null hypothesis.

3.2.2 Test 1: Discordant genotype frequency between batches

In samples drawn from the same ancestral population we would not expect differences in genotype frequencies between batches at the same marker beyond what would be expected by chance, given the available sample size and the frequency of the alternate allele. To detect deviation in genotype frequency of markers between batches we tested the null hypothesis that the genotype frequency was the same for a given marker between a single batch and the remaining batches combined. We used a Fisher's exact test on the 2x3 table of genotype counts (or 2x2 table for haploid markers). The majority of markers did not exhibit significant deviation in genotype frequency (p-value $< 3.15 \times 10^{-10}$, N=779,656). The remaining markers showed significant deviation in one or more batches (Figure 3). The marker quality control file contains the count of failed batches for each marker.

3.2.3 Test 2: Departure from Hardy-Weinberg equilibrium

Using PLINK (`--hardy` option), we conducted the test for departure from Hardy-Weinberg equilibrium (HWE) using the exact test [21] for each marker in each batch. Only diploid regions of the genome were

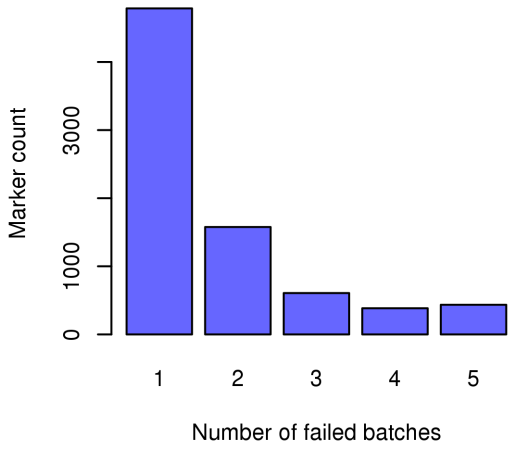

Figure 4: Count of markers that failed Hardy-Weinberg equilibrium test in one or more batches.

tested, including all autosomes, the pseudo-autosomal regions on the X chromosome, and females only on the sex-specific region of the X chromosome. There were 7,790 markers with a HWE p-value less than the multiple-testing corrected threshold of 3.15×10^{-10} in one or more batches (Figure 4), with relatively fewer markers that failed in more than one batch (N=3,003). The marker quality control file contains the count of failed batches for each marker.

3.2.4 Test 3: Discordance across control replicates

There were 4 control samples on each genotyping plate: Axiom Reference Genomic DNA 103, CEPH NA24385, CEPH NA12878 and CEPH NA10859. Using these control genotypes we determined the quality of the genotype calling for a specific marker as we expected the genotypes to be fully concordant for a control sample across all plates. For each marker and control sample we computed a discordance metric (d) as defined below:

$$d = 1 - \frac{max(n_{aa}, n_{ab}, n_{bb})}{n_{aa} + n_{ab} + n_{bb}}$$

where n_{aa}, n_{ab}, n_{bb} is the number of times the genotypes AA, AB, and BB are called for the individual at that marker. The distribution of genotype discordance (d) by control sample and batch is presented in Figure 5. There were 27,937 markers with control replicate discordance greater than 0.05 (that is less than 0.95 concordance) in one or more control samples. The marker quality control file contains the count of failed control samples for each marker.

3.2.5 Test 4: Genotype frequency discordance stratified between chromosomal sexes

We would not expect differences in genotype frequencies between chromosomally defined males and females for any marker except those on the Y chromosome (apart from loci associated with survival to recruitment age, that have sex-specific effects, e.g. X-linked recessive disorders). To detect deviation in

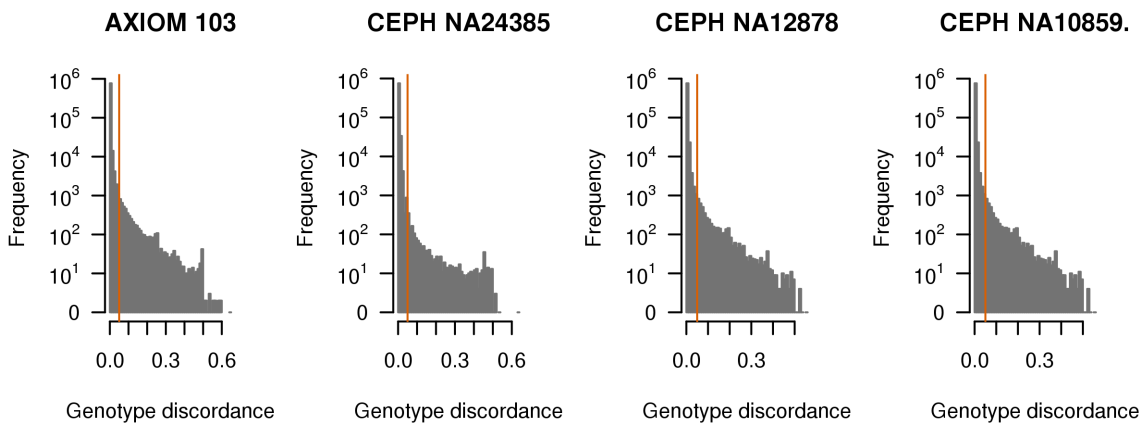

Figure 5: Frequency distribution of genotype discordance for control samples. Vertical line at genotype discordance threshold of 0.05.

genotype frequency of markers between the chromosomal sexes we tested the null hypothesis that genotype frequency is the same for a given marker between males and females. We used Fisher's exact test on the 2x3 table of genotype counts (or 2x2 table of allele counts for the sex-specific region of the X chromosome). There were 248 markers with discordant genotype counts or allele counts between batches with Fisher's p-value $< 3.15 \times 10^{-10}$ (Figure 6). These are labeled in the marker quality control file.

3.2.6 Flag low frequency SNPs and insertions/deletions

In addition to the above 4 tests, insertions/deletions and low-frequency SNPs (MAF < 0.005) were flagged as they may bias subsequent sample-based quality control (see section 4). There were 95,363 low-frequency SNPs (MAF < 0.005) and 15,616 indels.

3.3 Summary of results from marker-based tests

There were 37,706 markers that were flagged by one or more of the 4 tests. The effect of this quality analysis is depicted by comparing Figure 7 with Figure 8 where there is clear improvement in the concordance in minor allele frequency between batches after removal of these markers. Results for the above four tests are provided for all of the 794,409 markers and available within the marker quality control file. It is recommended to remove the 37,706 markers that fail one or more of the 4 quality control tests. In addition, we invite researchers to further exclude makers with low MAF and indels dependent on the analytical requirements. For instance, subsequent analyses in this document (see section 4 and section 5) have excluded markers that are within one or both of these categories.

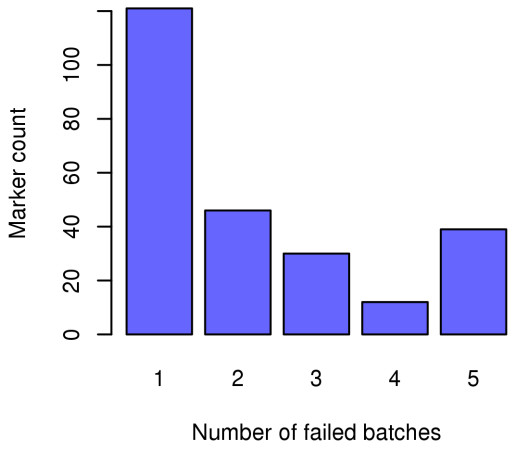

Figure 6: Count of markers that failed discordant genotype frequency test between males and females in one or more batches.

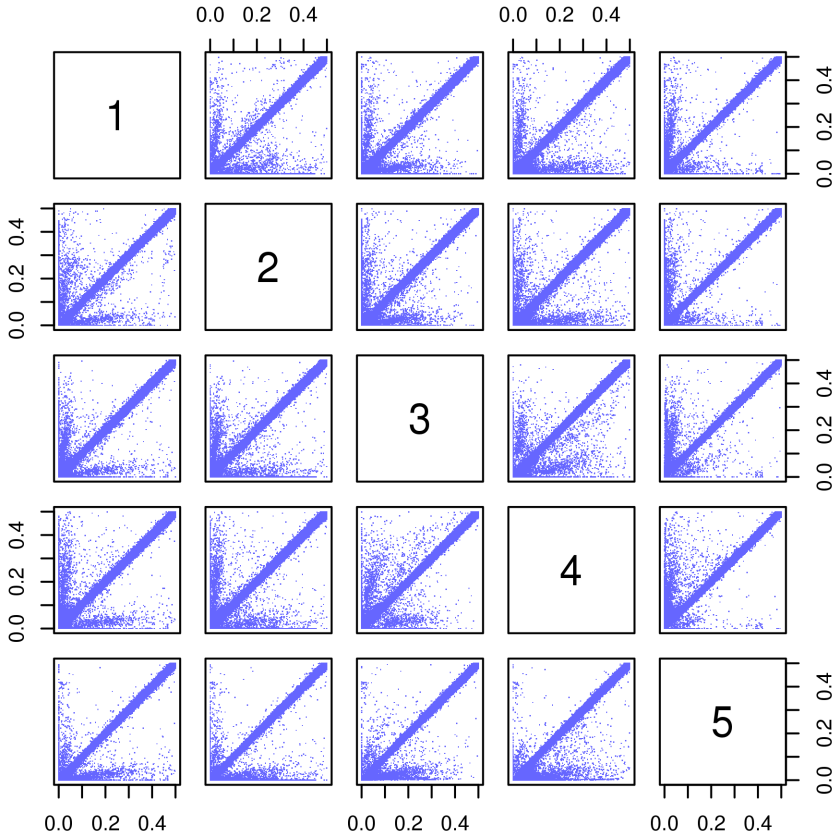

Figure 7: Pairwise plot of allele frequency of SNPs from genotype batch 1 to 4.

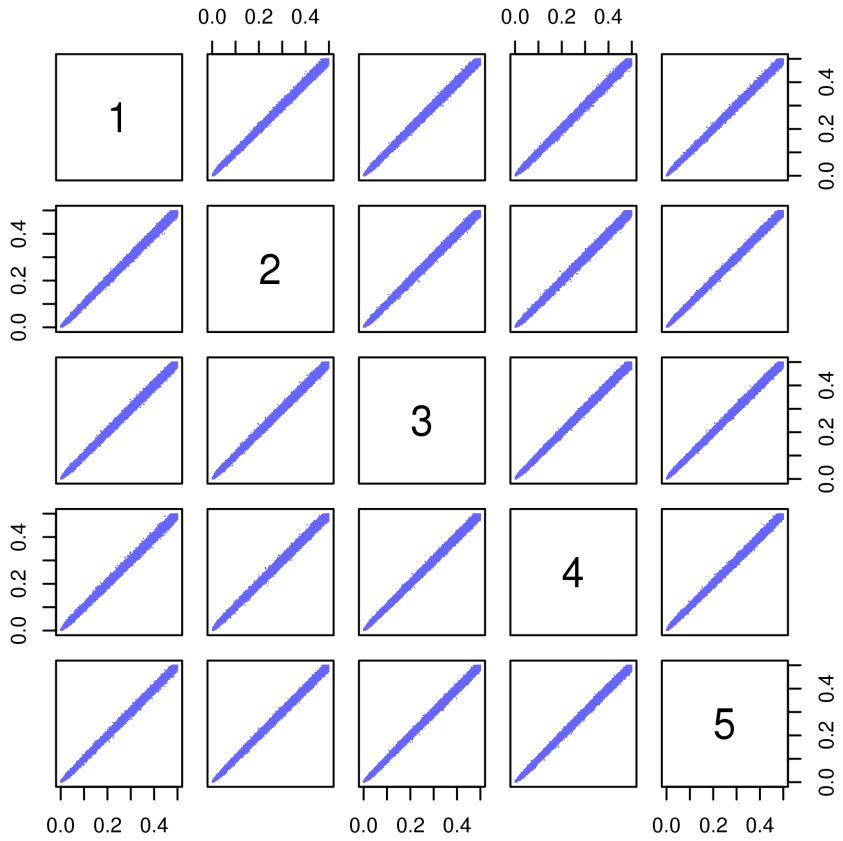

Figure 8: Pairwise plot of allele frequency of SNPs that pass all 4 tests from genotype batch 1 to 5.

4 Sample-based quality control

This sample-based quality control was intended to identify genotyped samples of low-quality, identify related individuals, and provide a genetic-based description of ancestry. We used a set of high quality SNPs to ensure that no bias was introduced due to genotype batch effect or other genotyping artifacts (see section 3.2). The purpose of this quality control analysis was to quantify the above mentioned properties of each genotyped sample, not to exclude samples from this data release. We thus encourage researchers using this data release to filter samples based on these properties, or devise their own sample quality control metrics that satisfy their research requirements.

4.1 Selection of markers for sample-based quality control

We selected among the 794,409 genetic markers based on the following criteria:

- Passed all 5 tests from marker-based quality control (see section 3.2) (removed 37,706 markers).
- Was not an insertion/deletion.
- Minor allele frequency > 0.01.
- Marker-wise missingness < 0.01.

The above filtering resulted in a total of 573,386 genetic markers. These genetic markers where then pruned to a set of 161,536 independent markers (PLINK option `--indep-pairwise 1000 kb 5 0.1`).

4.2 Determine familial relatedness

Familial relationships among CLSA participants were not recorded in the questionnaires or interviews. However, this information is essential for some epidemiological and most genomic analyses. Using the KING software program version 2.1.3 [10] we computed all pairwise kinship coefficients and noted all pairs with inferred relatedness of 3rd degree or closer (Table 2). KING uses both the analysis of identity-by-descent (IBD) segments as well as proportion of IBD and kinship coefficient to infer relatedness. Figure 9 depicts results from these two analyses and individuals are labeled to denote inferred relatedness. Individuals with an inferred relationship of 3rd degree or closer are labeled in the sample quality control file and the sample relatedness file.

Inferred Relationship	Count
Unrelated	123294
3rd degree	1066
Full Sibling	357
2nd degree	315
Parent Offspring	176
MZ Twin	1

Table 2: Count of kinship pairs per type of inferred relationship.

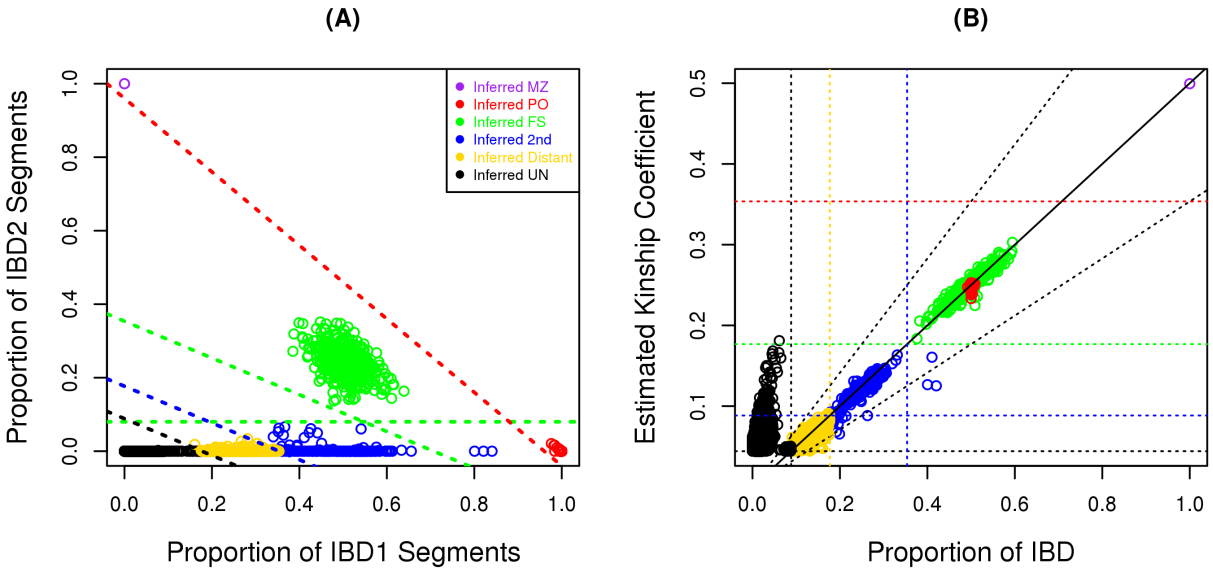

Figure 9: Inference of familial relatedness using KING. (A) Inference using IBD segments. (B) Inference using proportion IBD and kinship coefficient. Relationships in legend are abbreviated as: MZ=Monozygotic twin, PO=Parent/offspring, FS=Full sibling, 2nd=Second-degree relative, 3rd=Third-degree relative, Distant=Greater than 3rd degree relative, UN=Unrelated. Limits for inferring relationship type are indicated by dashed lines that are color-coded to match those listed in the legend.

4.3 Detecting outliers in heterozygosity and missing rates

Sample-wise genotype quality is typically assessed via the analysis of heterozygosity and genotype missingness, where extreme values in these measures may suggest low quality genotyping or cross-contamination of biological samples. We computed heterozygosity using PLINK (`--het` option), which implements the following formula for determining the proportion of non-missing genotypes that are heterozygous:

$$h = \frac{N_{nm} - N_{hom}}{N_{nm}}$$

where N_{nm} is the count of non-missing genotypes and N_{hom} is the count of homozygous genotypes. Sample-wise genotype missingness was computed using PLINK (`--miss` option). We observed 15 individuals with extreme values in heterozygosity and missingness (Figure 10). These outliers are identified in the sample quality metrics file. A genotyped sample's heterozygosity is partially dependent on population structure because the frequency of the alleles at one or more genetic markers may differ between populations. To illustrate this effect we first grouped self-reported ancestry into fewer categories (Table 3), with counts for each of these categories presented in Table 4. Visualization of sample-wise heterozygosity versus genotype missingness revealed that as expected heterozygosity is dependent on self-reported ancestry (Figure 10).

Self-reported Ancestry	Category
Arab	Arab
West Asian	Arab
Black	Black
Chinese	East Asian
Japanese	East Asian
Korean	East Asian
Latin American	Latino
Don't know	Other
Other	Other
Refused	Other
South Asian	South Asian
Filipino	South-East Asian
Southeast Asian	South-East Asian
White	White

Table 3: Self-reported ancestry and derived category.

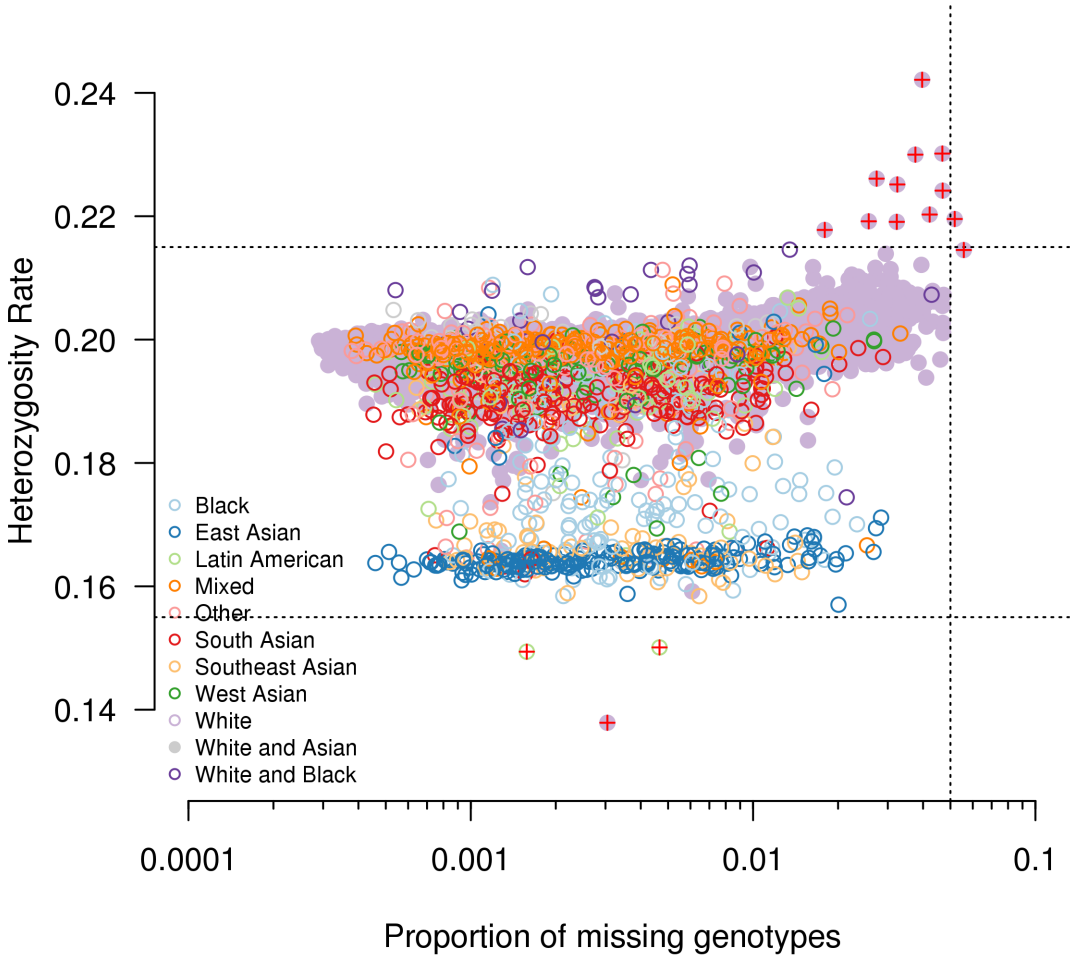

Figure 10: Sample-wise heterozygosity versus genotype missingness. Points are color coded according to ancestry category. Outliers are marked with a red plus sign.

Ancestry Category	Count
White	25172
Mixed	318
South Asian	223
East Asian	220
Other	172
Black	170
West Asian	105
Latin American	85
Southeast Asian	83
White and Asian	41
White and Black	33

Table 4: Count of genotyped individuals per ancestry category.

4.4 Population structure

Population structure was computed by projecting principal components from 1000 Genomes using the method developed by Abraham et al. (2017) [22]. This measure of population structure can be used to complement self-reported ancestry and is a common method for controlling for population stratification in genome-wide association studies [23, 24]. We first extracted the Affymetrix UK Biobank Axiom array markers from 1000 Genomes [20] and retained markers with the following criteria:

- MAF > 0.05.
- Hardy-Weinberg equilibrium (HWE) p-value $> 10^{-6}$.
- Single nucleotide substitutions with single character allele-codes (A, C, G, or T) (PLINK option `--snps-only just-acgt` option).
- Markers with unambiguous strand, that is exclude A/T or C/G markers.
- Retained linkage disequilibrium ($r^2 < 0.1$) pruned markers using window size of 1000 kb and step size of 5 markers (PLINK option `--indep-pairwise 1000 kb 5 0.1`).

This resulted in a set of 87,848 markers from which we computed principal component loadings on the 2504 individuals from 1000 Genomes phase 3. We then projected the 26622 CLSA individuals onto the principal components of the 1000 Genomes analysis. The pairwise comparison of the top 4 principal components are depicted in Figure 11.

4.5 Selection of the European ancestry subset

To reduce the effect of population structure on analyses such as GWAS it is recommended to use a subset of the population with relatively homogeneous ancestry. The majority of individuals in this genomic data release are of self-reported European ancestry (N=25,172, see section 4.4). We used genomic information and PCA analysis to determine a subset of individuals of homogeneous European ancestry, and refer to this subset as the ”CLSA European ancestry subset”.
To determine the CLSA European ancestry subset we clustered the top 4 principal components from the analysis of population structure in the previous section (see section 4.4) into 6 clusters. Visualization of these cluster alongside those from 1000 Genomes reveals a clear overlap of the largest cluster (cluster 4, N=24,658) with populations of European ancestry in 1000 Genomes. Moreover, this large cluster contains

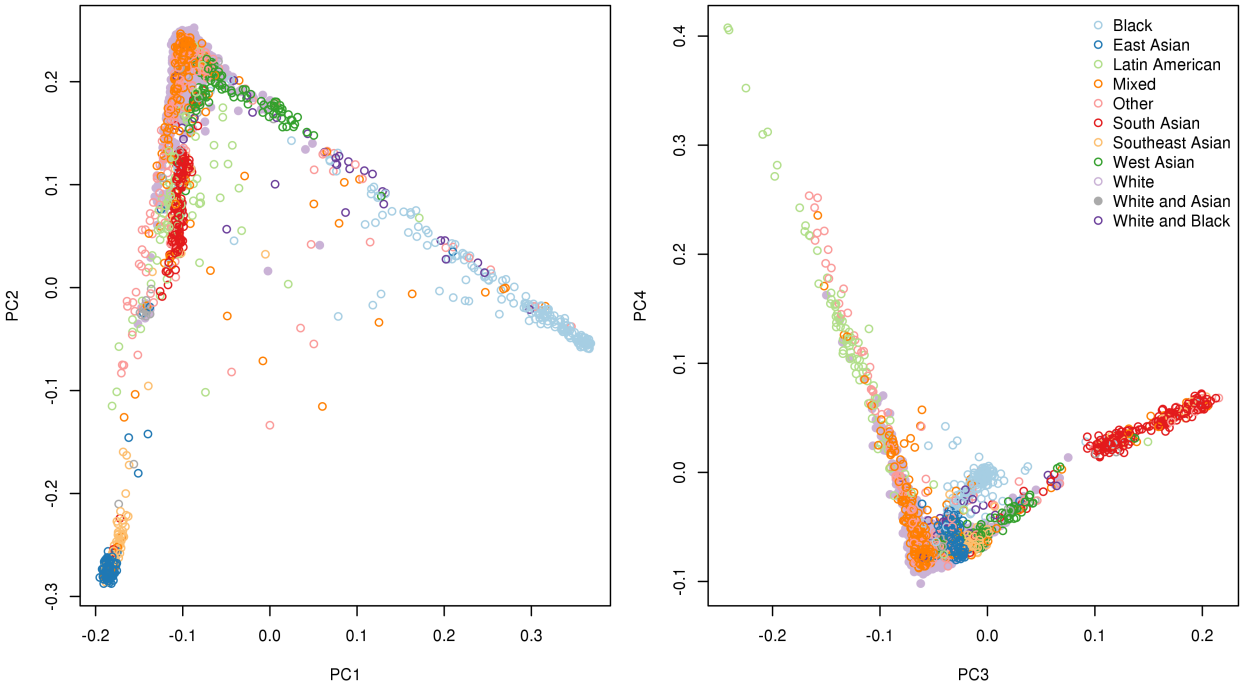

Figure 11: Population structure of CLSA genotyped participants. Pairwise plots of the top 4 principal components, with points color-coded according to self-reported ancestry groupings.

the vast majority of individuals in CLSA that self-report European ancestry (Table 5). The European ancestry subset has markedly reduced variance in the top principal components as compared to the entire CLSA cohort (Figure 13). The top 20 principal components of the PCA analysis are provided in the sample QC file accompanying this data release, as well as the top 10 principal components of the PCA analysis from the CLSA European ancestry subset.

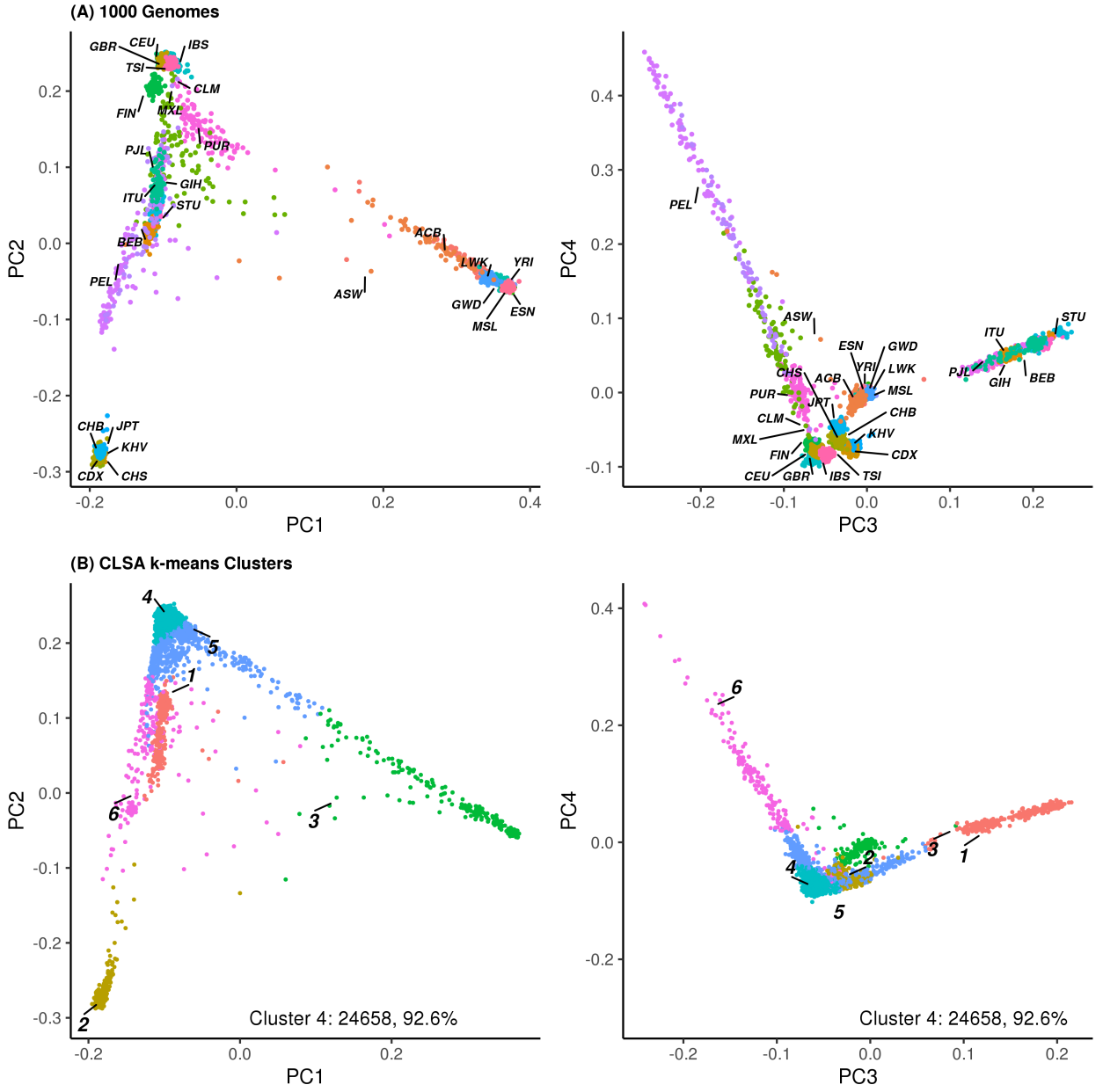

Figure 12: Determining the CLSA European ancestry subset. (A) Top 4 principal components from all 1000 Genomes populations labelled and coloured. (B) Top 4 principal components from CLSA color coded and labelled by cluster number.

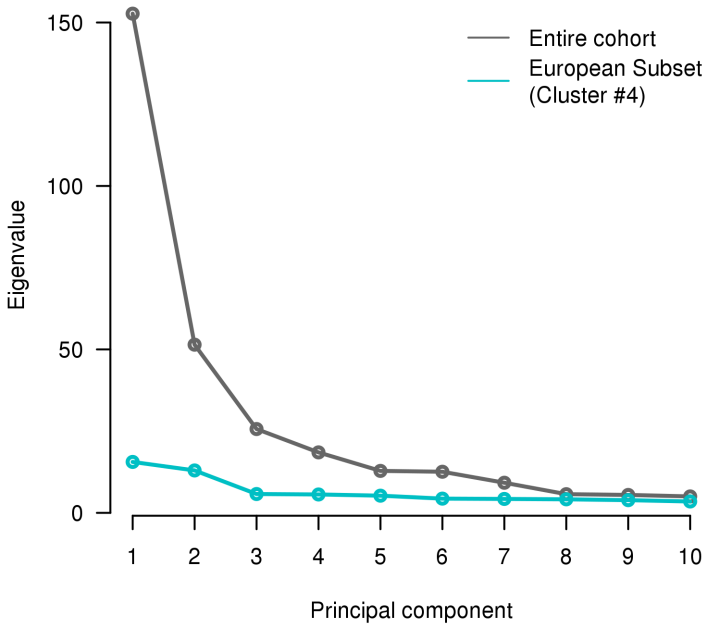

Figure 13: Eigenvalues for PCA analysis of the entire cohort (grey) and the European ancestry subset (cluster 4, light blue), demonstrating a reduction in genetic variance within the European ancestry subset.

	1	2	3	4	5	6
Black	7	0	156	0	7	0
East Asian	0	214	1	2	0	3
Latin American	1	0	1	2	9	72
Mixed	11	11	7	207	61	21
Other	11	5	8	54	53	41
South Asian	211	5	0	0	7	0
Southeast Asian	20	61	0	0	1	1
West Asian	4	0	1	2	98	0
White	7	2	0	24380	742	41
White and Asian	3	3	0	5	19	11
White and Black	2	0	11	3	17	0

Table 5: Count of individuals per self-reported ancestry and k-means cluster.

5 Genotype Imputation

5.1 Introduction

Genotype imputation is a computational method to predict marker genotypes that are not directly genotyped by an assay, such as a genotyping array. The imputation process uses a reference panel of densely whole genome sequenced individuals to predict genotypes in an array genotyped study sample for which only a subset of these genetic markers have been genotyped. This denser set of imputed markers increases the power of genetic association studies and enables fine-mapping of causal variants. For more information please refer to Fuchsberger et al. 2014 [25]. This imputed data set contains approximately 308 million genetic variants.

5.2 Marker filtering

As input to the imputation process we used the 26,622 CLSA participants that passed quality control (see section 4), and the set of 716,347 markers that satisfy the following criteria:

- Pass all marker 4 QC tests (see section 3.2).
- SNP-wise missingness < 0.05.
- Minor allele frequency > 0.0001.

Genotype imputation requires that the marker reference alleles match the human genome GRCh37 reference sequence. Using the `bcftools +fixref` plugin, this check removed 62,618 markers, resulting in a total marker count of 653,729 used as input into the imputation process.

5.3 Phasing, Imputation and Reference Panel

Phasing and imputation was conducted using the TOPMed reference panel [3] at the University of Michigan Imputation Service [26]. We used the TOPMed reference panel version r2, containing 97,256 reference samples at 308,107,085 genetic markers. We used this imputation service to pre-phase and impute the CLSA genotype data using EAGLE2 [27] and Minimac [25], respectively. Both autosomal and X chromosome variants were imputed. The imputation was carried out in two batches of 13,310 and 13,312 CLSA samples. Each batch also included the one of each 4 control samples (see section 1.4.2). The two batches where subsequently merged into a single dataset.

5.4 Imputation Performance

Imputation quality using the TOPMed reference panel was assessed using the marker-wise information measure (Rsq), and compared to the imputation using the Haplotype Reference Consortium reference panel containing 32,488 reference samples and 40.4 million genetic markers [28]. For each imputation data set, information measures for all SNPs on chromosome 22 were stratified into minor allele frequency (MAF) bins prior to comparison. Comparison of imputation quality between the two reference panels demonstrated that the TOPMed reference panel yielded overall higher imputation quality, likely due to the larger number of samples included in the reference panel (Figure 14).

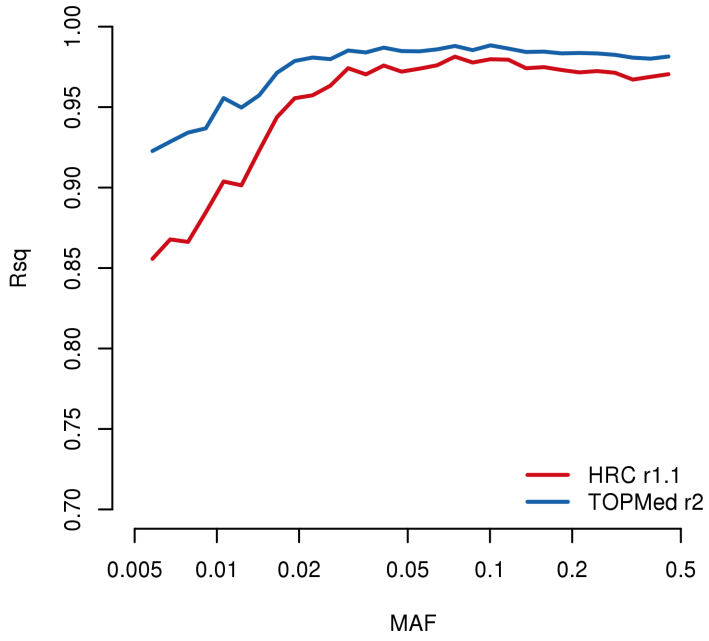

Figure 14: Imputation quality of the CLSA cohort using the TOPMed versus Haplotype Reference Consortium reference panels stratified by minor allele frequency bins (data shown is from chromosome 22).

References

[1] UK Biobank Axiom Array | UK Biobank. http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array, Apr 2018. [Online; accessed 10. Apr. 2018].

[2] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, Oct 2018.

[3] Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Quenna Wong, François Aguet, Christine Albert, Alvaro Alonso, Kristin G. Ardlie, Stella Aslibekyan, Paul L. Auer, John Barnard, R. Graham Barr, Lewis C. Becker, Rebecca L. Beer, Emelia J. Benjamin, Lawrence F. Bielak, John Blangero, Michael Boehnke, Donald W. Bowden, Jennifer A. Brody, Esteban G. Burchard, Brian E. Cade, James F. Casella, Brandon Chalazan, Yii-Der Ida Chen, Michael H. Cho, Seung Hoan Choi, Mina K. Chung, Clary B. Clish, Adolfo Correa, Joanne E. Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L. DeMeo, Susan K. Dutcher, Patrick T. Ellinor, Leslie S. Emery, Diane Fatkin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M. Fullerton, Soren Germer, Mark T. Gladwin, Daniel J. Gottlieb, Xiuqing Guo, Michael E. Hall, Jiang He, Nancy L. Heard-Costa, Susan R. Heckbert, Marguerite R. Irvin, Jill M. Johnsen, Andrew D. Johnson, Sharon L.R. Kardia, Tanika Kelly, Shannon Kelly, Eimear E. Kenny, Douglas P. Kiel, Robert Klemmer, Barbara A. Konkle, Charles Kooperberg, Anna Köttgen, Leslie A. Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth J.F. Loos, Lori Garman, Robert Gerszten, Steven A. Lubitz, Kathryn L. Lunetta, Angel C.Y. Mak, Ani Manichaikul, Alisa K. Manning, Rasika A. Mathias, David D. McManus, Stephen T. McGarvey, James B. Meigs, Deborah A. Meyers, Julie L. Mikulla, Mollie A. Minear, Braxton Mitchell, Sanghamitra Mohanty, May E. Montasser, Courtney Montgomery, Alanna C. Morrison, Joanne M. Murabito, Andrea Natale, Pradeep Natarajan, Sarah C. Nelson, Kari E. North, Jeffrey R. O'Connell, Nicholette D. Palmer, Nathan Pankratz, Gina M. Peloso, Patricia A. Peyser, Wendy S. Post, Bruce M. Psaty, D.C. Rao, Susan Redline, Alexander P. Reiner, Dan Roden, Jerome I. Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, Jeong-Sun Seo, Sudha Seshadri, Vivien A. Sheehan, M. Benjamin Shoemaker, Albert V. Smith, Nicholas L. Smith, Jennifer A. Smith, Nona Sotoodehnia, Adrienne M. Stilp, Weihong Tang, Kent D. Taylor, Marilyn Telen, Timothy A. Thornton, Russell P. Tracy, David J. Van Den Berg, Ramachandran S. Vasan, Karine A. Viaud-Martinez, Scott Vrieze, Daniel E Weeks, Bruce S. Weir, Scott T. Weiss, Lu-Chen Weng, Cristen J. Willer, Yingze Zhang, Xutong Zhao, Donna K. Arnett, Allison E. Ashley-Koch, Kathleen C. Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M. Rice, Stephen S. Rich, Edwin Silverman, Pankaj Qasba, Weiniu Gan, George J. Papanicolaou, Deborah A. Nickerson, Sharon R. Browning, Michael C. Zody, Sebastian Zöllner, James G. Wilson, L Adrienne Cupples, Cathy C. Laurie, Cashell E. Jaquish, Ryan D. Hernandez, Timothy D. O'Connor, and Gonçalo R. Abecasis. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *bioRxiv*, 2019.

[4] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. Genome-wide genetic data on 500,000 uk biobank participants. *bioRxiv*, 2017.

[5] Gavin Band and Jonathan Marchini. Bgen: a binary file format for imputed genotype and haplotype data. *bioRxiv*, 2018.

[6] QCTOOL v2. https://www.well.ox.ac.uk/~gav/qctool/index.html, Mar 2018. [Online; accessed 19. Jul. 2019].

[7] gavinband / bgen / wiki / bgenix — Bitbucket. https://bitbucket.org/gavinband/bgen/wiki/bgenix, Jul 2019. [Online; accessed 19. Jul. 2019].

[8] The BGEN format. https://www.well.ox.ac.uk/~gav/bgen_format, Apr 2019. [Online; accessed 19. Jul. 2019].

[9] Standard data input - PLINK 2.0. https://www.cog-genomics.org/plink/2.0/input#oxford, Aug 2020. [Online; accessed 6. Aug. 2020].

[10] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, Nov 2010.

[11] Affymetrix. UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory. http://biobank.ndph.ox.ac.uk/showcase/docs/affy_data_generation2017.pdf, 2017.

[12] Shaun Purcell and Christopher Chang. PLINK 1.9. https://www.cog-genomics.org/plink1.9, Apr 2018. [Online; accessed 27. Apr. 2018].

[13] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, Feb 2015.

[14] Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R. Nelson, and Gil McVean. Multi-Population Classical HLA Type Imputation. *PLoS Comput. Biol.*, 9(2):e1002877, Feb 2013.

[15] Alexander T. Dilthey, Pierre-Antoine Gourraud, Alexander J. Mentzer, Nezih Cereb, Zamin Iqbal, and Gil McVean. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLoS Comput. Biol.*, 12(10), Oct 2016.

[16] IPD-IMGT/HLA Database. https://www.ebi.ac.uk/cgi-bin/ipd/kir/fetch_cell.cgi?25906, Oct 2020. [Online; accessed 18. Nov. 2020].

[17] Chen-Shan Chin, Justin Wagner, Qiandong Zeng, Erik Garrison, Shilpa Garg, Arkarachai Fungtammasan, Mikko Rautiainen, Sergey Aganezov, Melanie Kirsche, Samantha Zarate, Michael C. Schatz, Chunlin Xiao, William J. Rowell, Charles Markello, Jesse Farek, Fritz J. Sedlazeck, Vikas Bansal, Byunggil Yoo, Neil Miller, Xin Zhou, Andrew Carroll, Alvaro Martinez Barrio, Marc Salit, Tobias Marschall, Alexander T. Dilthey, and Justin M. Zook. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.*, 11(4794):1–9, Sep 2020.

[18] Bastien Llamas, Giuseppe Narzisi, Valerie Schneider, Peter A. Audano, Evan Biederstedt, Lon Blauvelt, Peter Bradbury, Xian Chang, Chen-Shan Chin, Arkarachai Fungtammasan, Wayne E. Clarke, Alan Cleary, Jana Ebler, Jordan Eizenga, Jonas A. Sibbesen, Charles J. Markello, Erik Garrison, Shilpa Garg, Glenn Hickey, Gerard R. Lazo, Michael F. Lin, Medhat Mahmoud, Tobias Marschall, Ilia Minkin, Jean Monlong, Rajeeva L. Musunuri, Sagayamary Sagayaradj, Adam M. Novak, Mikko Rautiainen, Allison Regier, Fritz J. Sedlazeck, Jouni Siren, Yassine Souilmi, Justin Wagner, Travis Wrightsman, Toshiyuki T. Yokoyama, Qiandong Zeng, Justin M. Zook, Benedict Paten, and Ben Busby. A strategy for building and using a human reference pangenome. *F1000Research*, 8:1751, Oct 2019.

[19] William J. Rowell, Paul Peluso, Harting John, Yufeng Qian, Aaron Wenger, Richard Hall, and David R. Rank. Comprehensive variant detection in a human genome with PacBio high-fidelity reads. https://www.pacb.com/wp-content/uploads/Rowell-CSHLBioData-2018-Comprehensive-Variant-Detection-in-a-Human-Genome-with-PacBio-High-Fidelity-Reads.pdf, 2018. [Online; accessed 18. Nov. 2020].

[20] Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, Jul 2018. [Online; accessed 19. Jul. 2018].

[21] J. E. Wigginton, D. J. Cutler, and G. R. Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.*, 76(5):887–893, May 2005.

[22] Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, Sep 2017.

[23] David J. Balding. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7(10):781–791, Oct 2006.

[24] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, Jul 2006.

[25] Christian Fuchsberger, Gonalo R. Abecasis, and David A. Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, 10 2014.

[26] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, Emily Y. Chew, Shawn Levy, M K Mcgue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G. Iacono, Anand Swaroop, Laura J. Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonalo R. Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nature Genetics*, 48:1284–1287, 2016.

[27] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A. Reshef, Hilary K. Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R. Abecasis, Richard Durbin, and Alkes L. Price. Reference-based phasing using the Haplotype Reference Consortium panel. *bioRxiv*, page 052308, Jan 2016.

[28] The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, 48(10):1279–1283, Aug 2016.