BMJ Open

Cohort Profile: Genomic Data for 26,622 Individuals from the Canadian Longitudinal Study on Aging (CLSA)

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059021
Article Type:	Cohort profile
Date Submitted by the Author:	05-Nov-2021
Complete List of Authors:	Forgetta, Vince; Jewish General Hospital, Centre for Clinical Epidemiology Li, Rui; McGill University, Darmond-Zwaig, Corinne; McGill University Belisle, Alexandre; McGill University Balion, Cynthia; McMaster University, Pathology and Molecular Medicine Roshandel, Delnaz; The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning Wolfson, Christina; McGill University Lettre, Guillaume; Université de Montréal; Montreal Heart Institute Pare, Guillaume ; McMaster University Paterson, Andrew; Hospital for Sick Children, Griffith, Lauren; McMaster University, Department of Health Research Methods, Evidence, and Impact Verschoor, Chris; McMaster University, Lathrop, Mark; McGill University, Department of Human Genetics Kirkland, Susan ; Dalhousie University, Raina, Parminder; McMaster University, Clinical Epidemiology and Biostatistics Richards, Brent ; McGill University, Ragoussis, Jiannis; McGill University, Department of Human Genetics; McGill Genome Centre
Keywords:	GENETICS, EPIDEMIOLOGY, PUBLIC HEALTH, Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, SLEEP MEDICINE

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title: Cohort Profile: Genomic Data for 26,622 Individuals from the Canadian Longitudinal Study on Aging (CLSA)

Author List:

Vincenzo Forgetta[1†], Rui Li[2†], Corinne Darmond-Zwaig[2], Alexandre Belisle[2], Cynthia Balion[3], Delnaz Roshandel[4], Christina Wolfson[5], Guillaume Lettre[6], Guillaume Pare[3], Andrew D. Paterson[4,7,8], Lauren E. Griffith[9], Chris Verschoor[9], Mark Lathrop[2], Susan Kirkland[10], Parminder Raina[9‡], J. Brent Richards[1,5,11,12‡], and Jiannis Ragoussis[2,12,13‡]

1 Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montréal, QC, Canada,

2 McGill University Genome Centre, Department of Human Genetics, McGill University, Montréal, QC, Canada,

3 Hamilton Regional Laboratory Medicine Program, McMaster University, St. Joseph's Hospital St. Lukes Wing, Hamilton, ON, Canada,

4 Genetics & Genomic Biology, The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, ON, Canada,

5 Department of Medicine, & of Epidemiology and Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada,

6 Montréal Heart Institute and Université de Montréal, Montréal, QC, Canada,

7 Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada,

8 https://orcid.org/0000-0002-9169-118X

9 Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada,

10 Department of Community Health and Epidemiology, Division of Geriatric Medicine, Dalhousie University, Halifax, Nova Scotia, Canada,

11 Department of Twin Research and Genetic Epidemiology, King's College London, London, UK,

12 Department of Human Genetics, McGill University, Montréal, QC, Canada,

13 Department of Bioengineering, McGill University, Montréal, QC, Canada,

* Corresponding author. McGill Genome Centre, 740 Avenue Dr. Penfield, Montreal, Québec, Canada H3A 0G1. Email: ioannis.ragoussis@mcgill.ca.

† Joint first authors.

‡ Joint senior authors.

Keyword: CLSA, genome-wide genotyping, aging, HLA

Word count: 4,102

Abstract

Purpose: The Canadian Longitudinal Study on Aging (CLSA) Comprehensive cohort was established to provide unique opportunities in studying the genetic and environmental contributions to human health and disease in aging process. The aim of this study is to describe the genomic data included in CLSA.

Participants: A total of 26,622 individuals from CLSA baseline data collection on 51,338 men and women aged 45 to 85 recruited between 2010 and 2015 have undergone genome-wide genotyping of DNA samples collected from blood. Comprehensive quality control metrics were measured on genetic marker and sample-wise respectively. The genotypes were imputed to the TopMed reference panel. Sex chromosome abnormalities were identified by copy number profiling. The genotypes were imputed for classical HLA genes at two-field (four-digit).

Findings to date: Of the 26,622 genotyped participants, 24,655 (92.6%) were identified as having European ancestry. This genomic data can be linked to physical, lifestyle, medical, economic, environmental, and psychosocial factors collected longitudinally in CLSA. CLSA

genomic dataset has been used as a validation cohort to test the contribution of polygenic risk score to screen individuals with high fracture risk. It is also a valuable resource to directly identify common genetic variation associated with conditions related to complex traits. One study has employed CLSA genomic data in a large-scale GWAS and identified novel variants associated with sleep apnoea. Taking advantage of the comprehensive interview and physical information collected in CLSA, this genomic dataset has been linked to psychosocial factors to investigate both the independent and interactive effects on cardiovascular disease.

Future plans: The DNA methylation, metabolomic and proteomic data are being generated. Ongoing studies focus on elucidating the role of genetic factors in cognitive decline and cardiovascular diseases. This genomic data resource is available upon request through CLSA data access application process.

Strengths and limitations of this study

- The genomic data in Canadian Longitudinal Study on Aging (CLSA) Comprehensive cohort provides whole-genome genotyping data on 794,409 markers and whole-genome imputed data on approximately 308 million genetic variants.
- The UK Biobank array used for genotyping is enriched with known markers associated with multiple phenotypes. The comprehensive pharmacogenomic and inflammation markers may be of particular interest since DNA methylation, metabolomic and proteomic data are being generated by CLSA.
- The CLSA cohort has completed the baseline sample collection. It continues to follow up the participants on a wide spectrum of qualitative and quantitative variables. This facilitates the research on the effect of interplay between genetics and environmental factors on age-related diseases.
- Potential limitations may include the relatively lower genotyping coverage in participants with non-European ancestry and inadequate power to discover very rare predisposition

variants. Such limitations associated with this type of data can be overcome by imputation and meta-analysis.

Introduction

The global life expectancy increased dramatically through the past two hundred years. In such times, the make-up of Canadian population has changed unprecedentedly. From 1977 to 2017, the senior population, i.e., people aged 65 and older, grew from 2 million to 6.2 million, which equaled to nearly 17% of its population size. However, this number is still rapidly rising. It is anticipated that by 2036 there will be 10.2 million senior people in Canada. In another word, in every 4 Canadians, there will be one senior person.

Along with the expanded human life expectancy, the prevalence of age-related diseases is strikingly increasing. Aged people experience progressive decline in functional integrity and homeostasis. This process is accompanied by increased risk of neurodegeneration, cardiovascular disease and cancer among many other diseases, which have become the most common causes of decreased life quality and late-life mortality. It adds substantial burden to individual and social health care system inadvertently. Age-related diseases have highly complex nature. Both the genetic and environmental factors play an important role as well as the interaction between them. Therefore, understanding of the underlying mechanisms of aging is highly in demand for sustaining longer lives with reduced loss of healthy years.

Studies on short-lived model organisms provided insights on several key genetical regulators in hallmark aging pathways, however, the identification of biomarkers of age and age-related disease in human is more complicated [1]. Over the past decade, genetic epidemiology methods emerged to be a powerful tool. The genome-wide association studies (GWAS) uncovered tens of genes and genetic variations that may dominate the variability of aging outcomes among people [2]. They shed light on multi-trait variants associated with diseases. However, the genetic effects are usually relatively moderate and altered by lifestyle and other environmental

determinants. More work is needed to fully deconvolute the interplay between genetics and extrinsic influences. This effort will be benefited by larger sample size and linked information on proteomics and epigenetics.

Cohort description

The Canadian Longitudinal Study on Aging (CLSA) is a national long-term study that recruited 51,338 men and women, aged 45-85 years at enrolment between 2010 and 2015 for baseline data collection [3]. It 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. It is composed of two complementary cohorts: the Tracking cohort of 21,241 participants who are interviewed by telephone and the Comprehensive cohort of 30,097 participants who are interviewed in person and provide blood and urine samples. The participants in the Comprehensive cohort were randomly selected from within 25-50 km of 11 data collection sites in seven provinces. A total of 27,170 (90.3%) Comprehensive cohort participants provided blood samples at baseline. The Comprehensive Cohort samples have been used to produce whole genome genotyping data. The data were collected to understand, individually and in combination, the impact of genetic variation in both maintaining health and in the development of disease and disability as people age. In this release of the CLSA genomic data, 26,622 participants have been genotyped using the Affymetrix UK Biobank Axiom array [4]. Qualified researchers from any country can access these genomic and phenotypic data via a formal data and sample access procedure described on the CLSA Data Preview Portal.

Patient and public involvement

Patients or public were not involved in the development of the research question and study design or conducting the present study.

Data collected:

Sample storage and DNA extraction

The CLSA protocol was reviewed and approved by 13 research ethics boards across Canada. All participants provided written informed consent [5]. The biological samples were collected at the Data Collection Site and de-identified. Whole blood buffy coats were isolated from peripheral blood drawn and the plasma layer was removed. Samples were immediately moved to -80°C storage, and transferred to LN_2 storage at the CLSA Biorepository and Bioanalysis Centre 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). 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.

Genotyping and calling

Each plate genotyped contained 92 CLSA DNA samples and 4 controls, one male control as the Affymetrix Reference Genomic DNA 103 (Catalog# 900421)) or Personal Genome Project sample huAA53E0 (Coriell Cell Repositories, catalog # NA24385), two female controls as the CEPH control 1463-02 (Coriell Cell Repositories, catalog # NA12878) or the CEPH control 1347-2 (Coriell Cell Repositories, catalog # NA10859), and a deionized water negative control. The Affymetrix protocol (Axiom 2.0 Assay Automated workflow on Affymetrix NIMBUS) was followed. Samples were hybridized to UK Biobank arrays (ThermoFisher Catalog #902502), the same array that was used to genotype ~450,000 individuals in the UK Biobank cohort [6]. Axiom Array plates were processed on the Affymetrix GeneTitan Multi-Channel Instrument. For first pass quality control (QC), batches of 8 plates were analyzed using the Sample QC workflow

of the AxiomTM Analysis Suite 2.0 software where a subset of 20,000 reliable probes were used to determine Dish QC (the measure of the resolution of the AT and GC signal contrast) and sample QC. The reliable probes are autosomal, previously wet-lab tested, working probe sets with two array features per probe set.

Genotyping quality control and removal of duplicate genotyped participants

Genotyping was undertaken in separate batches of approximately 5,000 samples each using AxiomTM Analysis Suite 2.0, similar to UK Biobank genotyping QC documentation [4]. Genotype calling resulted in 27,010 successfully genotyped DNA samples. An inclusion list containing 794,409 genetic variants was used [6], as well as the following QC parameters for selecting samples passing to further analysis: Dish QC ≥ 0.82 on sample level, and average QC call rate of passing samples on a plate (plate QC call rate) ≥ 95%, percentage of passing samples ≥ 70%, and average call rate for passing samples ≥ 95% on plate-level. Duplicate genotyped participants were detected by KING version 2.1.3 [7] and the sample with higher genotype missingness was removed. This resulted in 26,622 successfully genotyped participants.

Sex chromosome composition

Distribution of F estimates on the X chromosome showed a gap between 0.4 and 0.8 (Supplementary Figure S1). Using this threshold, we obtained X chromosome number using PLINK version 1.90b4.4 [8 9]. F estimates for the 48 individuals with sex discrepancies between self-reported sex and X chromosome composition (Table 1) are listed in Supplementary Table S1. All subsequent analyses in this paper will use X chromosome number to define sex.

Genetic marker-based quality control

This 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.

The above tests require a population with relatively homogenous ancestry. Given this, we determined the largest subset of ancestrally homogeneous participants via K-means clustering of projected principal components from 414 individuals across 4 populations (Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT) and Yoruba in Ibadan, Nigeria (YRI)) from 1000 Genomes Phase 3. The largest cluster across all genotype batches overlapped the CEU population, and included a total of 24,361 individuals, or 92% of the entire genotyped cohort (N=26,622) (Supplementary Figure S2).

We then 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 0.001 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. Single nucleotide polymorphisms (SNPs) that failed the tested QC parameters are flagged within the marker quality table provided with the data release. We thus invite researchers to filter markers based on these properties or devise their own quality control metrics that satisfy their research requirements.

Discordant genotype frequency between batches

To detect deviation in genotype frequency of markers between batches, we used a Fisher's exact test on the 2x3 table of genotype counts (or 2x2 table for haploid markers). The vast majority of markers did not exhibit significant deviation in genotype frequency (779,656, 98.1% of total).

Departure from Hardy-Weinberg equilibrium

We conducted the test for departure from HWE using the exact test [10]. There were 7,790 markers with an HWE p-value $< 3.5 \times 10^{-10}$.

Discordance across control replicates

There were 3 positive control samples on each genotyping plate: a male control (Affymetrix CTL1 103 or Personal Genome Project participant huAA53E0), and one of two female controls (CEPH 1463-02 or CEPH 1347-02) in duplicate. For each marker and control sample we computed a discordance metric (d) defined as 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. There were 27,937 markers with control replicate discordance greater than 0.05 (i.e. concordance < 0.95).

Sex genotype frequency discordance

To detect deviation in genotype frequency of markers between sexes, we used Fisher's exact test on the 2x3 table of genotype counts for autosomal SNPs (or 2x2 table of allele counts for the sex-specific regions of the X chromosome). There were 248 markers with discordant genotype counts or allele counts between sexes with p-value < 3.5×10^{-10}.

Summary of results from marker-based tests

There were 37,706 SNPs that were flagged by one or more of the 4 tests. They are labeled in the marker quality control file accompanying this data release. The effect of this quality analysis is depicted by comparing Supplementary Figure S3 with Figure 1 where there is clear improvement in the concordance in minor allele frequency between batches after removal of these markers. We recommend to remove these markers, but have maintained these markers in the dataset so that researchers have access to all data. In addition, 15,616 insertions/deletions(indels) and 95,363 low-frequency SNPs with minor allele frequency (MAF) < 0.005 were flagged as they may bias subsequent sample-based quality control.

Sample-based quality control

This sample-based quality control was intended to identify genotyped samples of low-quality, related individuals, and provide a genetic-based description of ancestry. We thus encourage

researchers using this information included in the data release to filter samples or devise their own sample quality control metrics that satisfy their research requirements.

We selected the SNP markers that passed all 4 tests from marker-based quality control with MAF > 0.01 and marker-wise missingness < 0.01 resulting in a total of 573,386 markers. The software program PLINK was used to LD- prune these markers to a subset of 161,536 independent markers that were used for the following sample-wise assessments. The pruning was done on window size of 5000 kb with pairwise r^2 threshold as 0.1 and the number of variants to shift the window as 5.

Familial relatedness

Familial relationships among CLSA participants were not recorded in the questionnaires or interviews. However, this information is essential for some epidemiological and genomic analyses. Using the KING software [7] we computed all pairwise kinship coefficients and noted all pairs with inferred relatedness of 3rd degree or closer using autosomal SNPs (Table 2, Supplementary Figure S4). Individuals with an inferred relationship of 3rd degree or closer are labeled in the database.

Detection of outliers in heterozygosity and missing rates

Since extreme values in sample-wise heterozygosity and missingness may suggest low quality genotyping or cross-contamination of biological samples, we detected outliers by using PLINK (Supplementary Figure S5). As expected, because the allele frequencies differ between populations, we observed that heterozygosity was dependent on self-reported cultural background.

Population structure

Population structure was computed by principal component analysis (PCA) [11] to complement self-reported ancestry and control for population stratification in GWAS [12 13]. The top 20 principal components were computed using a high-quality subset of unrelated individuals by removing

individuals classified as outliers in heterozygosity and missingness, and any individual with a relation of 3rd degree or less.

Selection of 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). We combined self-reported ancestry with genomic information and PCA analysis to identify a subset of self-reported European individuals with relatively homogenous 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 into 6 clusters. Visualization of these clusters alongside those from 1000 Genomes reveals a clear overlap of the largest cluster (cluster 4, N=24,655) with populations of European ancestry in 1000 Genomes (Figure 2). Moreover, this largest cluster contains the vast majority of individuals in CLSA that self-report European ancestry (Table 3, Supplementary Table S2). The European ancestry subset has markedly reduced variance in the top principal components as compared to the entire CLSA cohort (Supplementary Figure S6). 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.

Detection of copy number abnormalities associated with disease

Sex chromosome abnormalities

The sex was called by both Affymetrix Axiom™ Analysis Suite 2.0 and PLINK. Affymetrix uses the ratio of mean signal values of non-polymorphic probes separately on the X and Y chromosomes to calculate sex. PLINK determines sex by using only X chromosome inbreeding coefficient (F estimates). When a subject has sex chromosomal abnormalities such as Turner

syndrome (45, X), Affymetrix will call them female but PLINK will call them male. Similarly, when a subject has Klinefelter Syndrome (47, XXY), Affymetrix will call the subject male but PLINK will call them female. We use this discordance information combined with copy number profiling to identify chromosomal abnormalities in CLSA participants.

To correct the miscalling of males by stringent Affymetrix default threshold, the intensity data of chromosome X and Y markers from all UK Biobank samples were used as a training data set to generate a Support Vector Machine (SVM) model. This SVM model was applied to CLSA samples to recall the vast majority of miscalled samples (331 out of 359). However, the SVM approach as aforementioned could not be applied to PLINK sex calling since the sex calling in UK Biobank data was already corrected. Alternatively, an empirical threshold was used to recall most (140 out of 175) of the samples miscalled by PLINK through setting X chromosome F estimate < 0.3 as female and > 0.8 as male. We used a relatively more stringent threshold of F estimate because high F estimates may indicate mosaic chromosomal abnormalities such as mosaic deletion. Finally, we used Axiom CNV Summary Tool to calculate log2 ratio and B allele frequency (BAF, which is in fact the within person ratio of B/B+A intensity at each SNP) for both X and Y chromosomes from the genotyping data. The log2 ratio and BAF were used to identify sex chromosomal abnormalities compared to normal male and female (Figure 3 (A-B)).

As a result, we detected 63 participants with discordance between self-reported sex and Affymetrix and/or PLINK sex calling (Supplementary Table S2), then we examined their CNV to identify them as one of four scenarios, sex chromosomal aneuploidy (11 subjects), mosaic sex chromosomal aneuploidy (15 subjects), low heterozygosity on the X chromosome (14 subjects), discordance between X chromosome number and self-reported sex without sex chromosomal aneuploidy (23 subjects). Briefly, we identified all 5 participants with self-reported sex chromosomal abnormalities including 1 mosaic Turner syndrome patient (45,X/46,XY) (scenarios 1 and 2). We identified all 48 participants with sex discordance as in abovementioned sex check. For the 23 participants who had discordance with both Affymetrix

and PLINK calling, CNV analysis confirmed the sex chromosome composition (scenario 4). In addition, for participants with no self-reported sex, Affymetrix/PLINK calling and CNV analysis are concordant to call sex. Besides the validated self-reported sex chromosomal abnormalities, we identified 4 participants with Klinefelter syndrome (47,XXY) and 3 with Turner Syndrome (45,X) (scenario 1) (Figure 3 (C-D)). In total, we found 3 participants with 45,X/46,XX mosaicism, and 11 participants with 45,X/46,XY mosaicism including 1 with self-reported Turner syndrome (45,X/46,XY) (Figure 3 (E-F)). Additionally, individuals with low heterozygosity on chromosome X could be a result of inbreeding (Supplementary Figure S7).

Charcot-Marie-Tooth Disease

Charcot-Marie-Tooth disease (CMT) is one of the most common inherited neurological disorders. It is mostly caused by duplication at 17p12 where *PMP22* is located (CMT1A and CMT1E; OMIM: # 118220; # 118300). In this release of CLSA genomic data, there are 9 CLSA participants who self-reported as having CMT. We examined their CNVs and found that 4 participants have duplication at *PMP22* (Supplementary Figure S8), and 1 participant has deletion at *PMP22* (Supplementary Figure S8). The other 4 subjects did not have CNVs detected at *PMP22*.

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. Calling accuracy was 100% across the loci (Supplementary Table S3). The imputation accuracy of genotyped CLSA participants was estimated by using the replicated samples. The validation rate is 100% for all the replicates.

Imputation to the TopMed reference panel

Genotype imputation is a computational method to predict marker genotypes that are not directly genotyped by an assay, such as genotyping array. The imputation process uses a reference panel of sequenced individuals to predict genotypes in a study sample for which only a subset of these genetic markers has been genotyped [15]. As input to the imputation process, we used the 26,622 CLSA participants that passed quality control, and the set of 653,729 markers that pass all marker QC tests, with SNP-wise missingness < 0.05, MAF > 0.0001 and have alleles that match the human genome GRCh37 reference sequence.

Phasing and imputation were conducted using the TOPMed reference panel [16] at the University of Michigan Imputation Service [17]. 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 [18] and Minimac [15], 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 3 control samples. The two batches where subsequently merged into a single dataset.

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 [19]. For each imputation data set, information measures for all SNPs on chromosome 22 were stratified into 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 (Supplementary Figure S9).

Findings to date

This data resource has been used in three completed and several ongoing studies. In a study to investigate the contribution of polygenic risk score (GRS) to screening for fracture risk [20], the CLSA genomic data was linked to the participants' physical examinations. It was the largest cohort included in this facture risk study for testing, which enabled the researchers to understand the performance of GRS particularly in old-aged individuals. It was found that the genetic pre-screening could reduce the number of further assessments to identify individuals at high risk of osteoporotic fractures. In another study on cardiovascular disease [21], the investigators evaluated the independent effects and interactions of multiscale risk factors by taking advantage of combined genomic and psychosocial information collected in CLSA cohort. In addition, the CLSA dataset provides opportunities to study other conditions related to complex diseases. It was employed by a large scale GWAS on sleep apnoea which was associated with cardiovascular disease and glaucoma. The authors revealed robust novel associations between 30 genes and this condition, and substantial molecular overlap with other complex traits [22]. For further publications please consult https://www.clsa-elcv.ca/stay-informed/publications.

Strengths and limitations

The CLSA genomic data are a unique resource nested in a large-scale, longitudinal study profiling aging population in Canada. The genotyping array is enriched with known markers associated with multiple phenotypes. However, the UK Biobank array may have relatively lower coverage in participants with non-European ancestry. The sample size may be small to identify very rare variants. In spite of these limitations, CLSA cohort includes deep and extensive phenotyping and planned linkage to health administrative databases. This data resource will facilitate the research on complex relationship between human genomic variants and a wide spectrum of environmental, lifestyle, and medical factors. The comprehensive pharmacogenomic and inflammation markers among other disease-associated variants may be

of particular interest since DNA methylation, metabolomic and proteomic data are being generated.

Collaboration

The genomic data from the CLSA Comprehensive cohort are accessible via the CLSA Data Access process (https://www.clsa-elcv.ca/data-access). The list of phenotypic variables can be browsed via the CLSA Data Preview Portal (https://datapreview.clsa-elcv.ca/). To be informed of the potential overlapping research topics, prospective data users are encouraged to consult the approved project summaries catalogued on the CLSA website (http://www.clsa-elcv.ca/researchers/approved-project-summaries). Given that this genomic data resource is released in 2018, we calculated the proportion of data requests including genomic data since 2018. At the time of writing, 17% of approved projects requested genetic data for their studies. The directly genotyped data are provided in binary PLINK format. It is recommended to use PLINK to manipulate these files (https://www.cog-genomics.org/plink/1.9/). The imputed genotyped data are provided in binary BGEN version 1.2 format using 8-bit encoding. It is recommended to use *qctool* version 2 or *bgenix* to manipulate this data type. The HLA imputation file is a plaintext file containing information pertaining to the imputation of classical human leukocyte antigen alleles from SNP genotypes.

All studies using CLSA genetic data resource are requested to give full acknowledgement to CLSA in their publications following instructions in *Publication and Promotion Policy for CLSA Data Users* on https:///www.clsa-elcv.ca.

Funding

Funding for CLSA is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under grant reference: LSA 94473 and the Canada Foundation for

Innovation. The work was also supported by Genome Canada Technology Platform #12505 and CFI#33408.

Author contributions

V.F. and R.L. conducted data analyses and drafted the manuscript, C.D-Z. and A.B. generated data, C.B., D.R., C.W., G.L., G.P., A.D.P., L.E.G., C.V., M.L., S.K., P.R., J.B.R., and J.R developed the concept and study design. All authors revised the manuscript critically for important intellectual content and approved the final version to be published.

Competing interests: None declared.

Reference:

1. Singh PP, Demmitt BA, Nath RD, et al. The Genetics of Aging: A Vertebrate Perspective. *Cell* 2019;177(1):200-20. doi: 10.1016/j.cell.2019.02.038 [published Online First: 2019/03/23]
2. Melzer D, Pilling LC, Ferrucci L. The genetics of human ageing. *Nat Rev Genet* 2020;21(2):88-101. doi: 10.1038/s41576-019-0183-6 [published Online First: 2019/11/07]
3. Raina P, Wolfson C, Kirkland S, et al. Cohort Profile: The Canadian Longitudinal Study on Aging (CLSA). *Int J Epidemiol* 2019;48(6):1752-53j. doi: 10.1093/ije/dyz173 [published Online First: 2019/10/22]
4. Affymetrix. UKB WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory. 2017. http://biobank.ndph.ox.ac.uk/showcase/docs/affy_data_generation2017.pdf.
5. Raina PS, Wolfson C, Kirkland SA, et al. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009;28(3):221-9. doi: 10.1017/S0714980809990055 [published Online First: 2009/10/29]
6. UK Biobank Axiom Array | UK Biobank [Available from: http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ accessed 10. Apr. 2018.
7. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26(22):2867-73. doi: 10.1093/bioinformatics/btq559 [published Online First: 2010/10/12]
8. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7. doi: 10.1186/s13742-015-0047-8 [published Online First: 2015/02/28]
9. Chang SPaC. PLINK 1.9 [Available from: https://www.cog-genomics.org/plink1.9 accessed 27. Apr 2018.
10. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005;76(5):887-93. doi: 10.1086/429864 [published Online First: 2005/03/25]
11. Galinsky KJ, Bhatia G, Loh PR, et al. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 2016;98(3):456-72. doi: 10.1016/j.ajhg.2015.12.022 [published Online First: 2016/03/01]
12. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7(10):781-91. doi: 10.1038/nrg1916 [published Online First: 2006/09/20]
13. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38(8):904-9. doi: 10.1038/ng1847 [published Online First: 2006/07/25]
14. Dilthey A, Leslie S, Moutsianas L, et al. Multi-population classical HLA type imputation. *PLoS Comput Biol* 2013;9(2):e1002877. doi: 10.1371/journal.pcbi.1002877 [published Online First: 2013/03/06]

15. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* 2015;31(5):782-4. doi: 10.1093/bioinformatics/btu704 [published Online First: 2014/10/24]
16. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;590(7845):290-99. doi: 10.1038/s41586-021-03205-y [published Online First: 2021/02/12]
17. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-87. doi: 10.1038/ng.3656 [published Online First: 2016/08/30]
18. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48(11):1443-48. doi: 10.1038/ng.3679 [published Online First: 2016/10/28]
19. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-83. doi: 10.1038/ng.3643 [published Online First: 2016/08/23]
20. Forgetta V, Keller-Baruch J, Forest M, et al. Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study. *PLoS Med* 2020;17(7):e1003152. doi: 10.1371/journal.pmed.1003152 [published Online First: 2020/07/03]
21. Menniti G, Paquet C, Han HY, et al. Multiscale Risk Factors of Cardiovascular Disease: CLSA Analysis of Genetic and Psychosocial Factors. *Front Cardiovasc Med* 2021;8:599671. doi: 10.3389/fcvm.2021.599671 [published Online First: 2021/04/03]
22. Campos AI, Ingold N, Huang Y, et al. Genome-wide analyses in 1,987,836 participants identify 39 genetic loci associated with sleep apnoea. *medRxiv* 2020:2020.09.29.20199893. doi: 10.1101/2020.09.29.20199893

Table1: Count of CLSA genotyped participants by self-reported gender and sex chromosome composition

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

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

Inferred Relationship	Count
Monozygotic twin	1
Full sibling	357
Parent/offspring	176
2nd degree	315
3rd degree	1066
Unrelated	123294

Table 3: Count of CLSA genotyped participants per self-reported ancestry and k-means cluster

Self-reported ancestry[a]	k-means cluster					
	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

[a]The details of grouping self-reported cultural and racial category into fewer groups are in Supplementary Table S2

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

The SNPs are considered as passed if they have nonsignificant p-value (Fisher's $p > 3.5 \times 10^{-10}$) below the multiple testing corrected threshold for the respective test on discordant genotype frequency between batch, departure from HWE, discordance between the positive control replicates and on discordant genotype frequency between male and female.

Figure 2: Determining the CLSA European ancestry subset.

(A) Top 4 principal components from all 1000 Genomes populations labelled and coloured. Population code refers to https://www.internationalgenome.org/category/population/. (B) Top 4 principal components from CLSA color coded and labelled by cluster number.

Figure 3: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X and Y are shown for sex chromosome abnormalities.

(A) In 46,XY, the BAF is either 0 or 1 and the expected log2 Ratio is less than 0 on chromosome X. However, in the pseudoautosomal region (PAR) and the chrY11.2/chrXq21.3 homology block, there are heterozygous calls in male shown as BAF of 0.5. The red line shows the lowess curve for log2 Ratio. The BAF is either 0 or 1 and the expected log2 Ratio is 0 on chromosome Y. (B) In 46,XX, the BAF is either 0 (AA), ½ (AB) or 1 (BB) and the expected Log2 Ratio is 0 on chromosome X as in a normal diploid cell. The BAF is between 0 and 1, and Log2 Ratio is less than 0 on chromosome Y. (C) For Klinefelter syndrome (47,XXY), log2 ratio is around 0 on chromosome X which indicates ploidy as 2N. Compared to 46,XY, there is relatively lower peaks of log2 ratio at PAR and chrX21.3/chrY11.2 homology block region. And BAF of heterozygous calls at PAR and chrX21.3/chrY11.2 homology block region shifted from 0.5 to intermediate values. They both indicated an extra copy of chromosome X. Chromosome Y intensity profile showed clear male pattern. (D) For Turner syndrome (45,X), on chromosome X, log2 ratio is below 0 and there is no BAF bands of 0.5, which indicates one copy loss. Chromosome Y intensity profile showed clear female pattern. (E) For 45,X/46,XX mosaicism, on chromosome X, there is a relatively smaller decrease of log2 ratio compared to 1 copy of chromosome X as in male. The BAF of heterozygous calls on chromosome X is split to intermediate values. They both indicate that the sample is mosaic for deletion of chromosome X. Chromosome Y intensity profile showed clear female pattern. (F) For 45,X/46,XY mosaicism, the log2 ratio less than 0 and no BAF 0.5 band on chromosome X indicates one copy. The log2 ratio shifts to below 0 and BAF values between 0 and 1 on chromosome Y indicates chromosome loss. However, the intermediate BAF values close to 0 or 1 at PAR and chrX21.3/chrY11.2 homology block region indicates the loss of chromosome Y is existed in a larger proportion of cells.

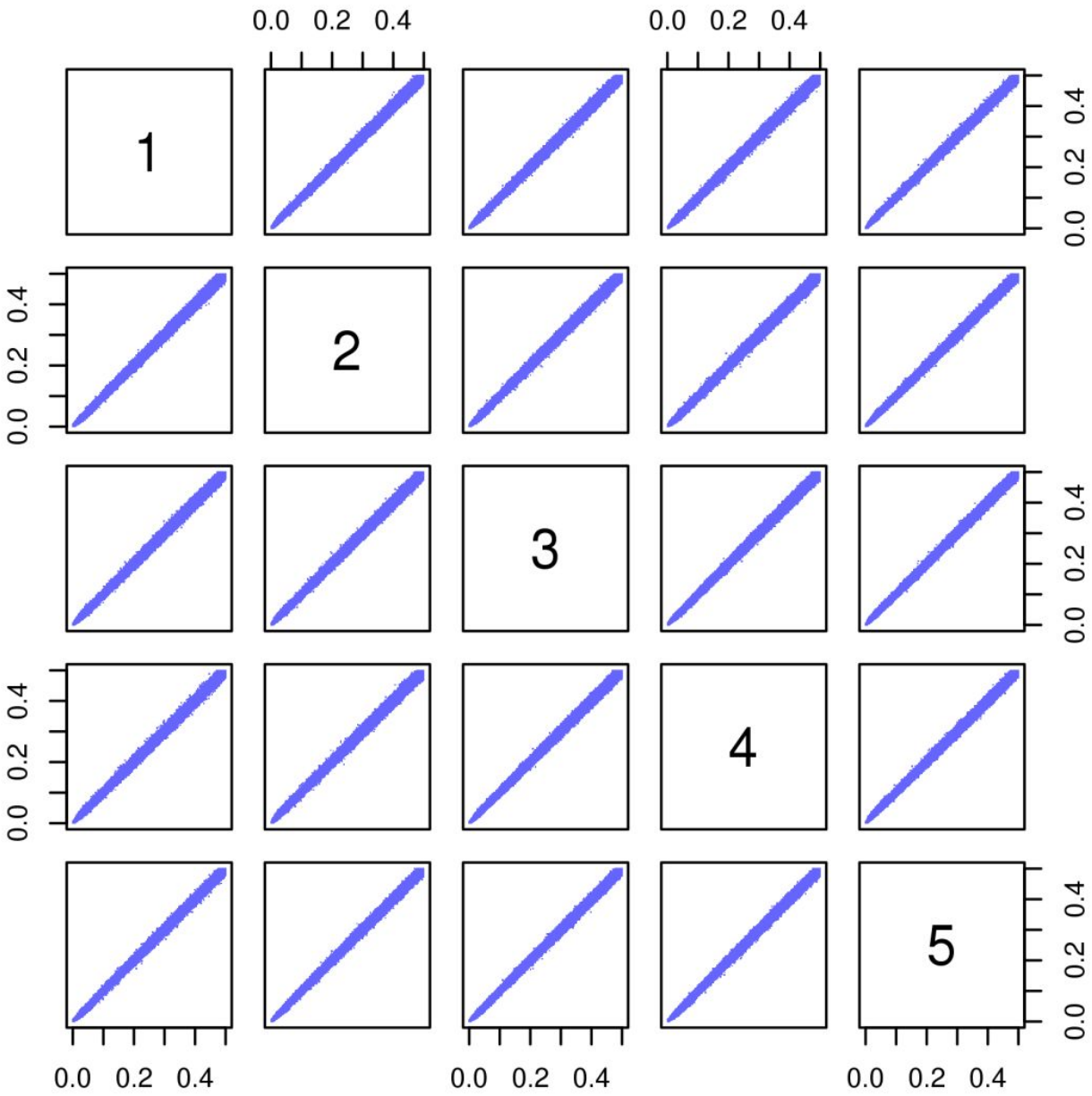


Figure 1: Pairwise plot of allele frequency of SNPs that pass all 4 tests from genotype batch 1 to 5.

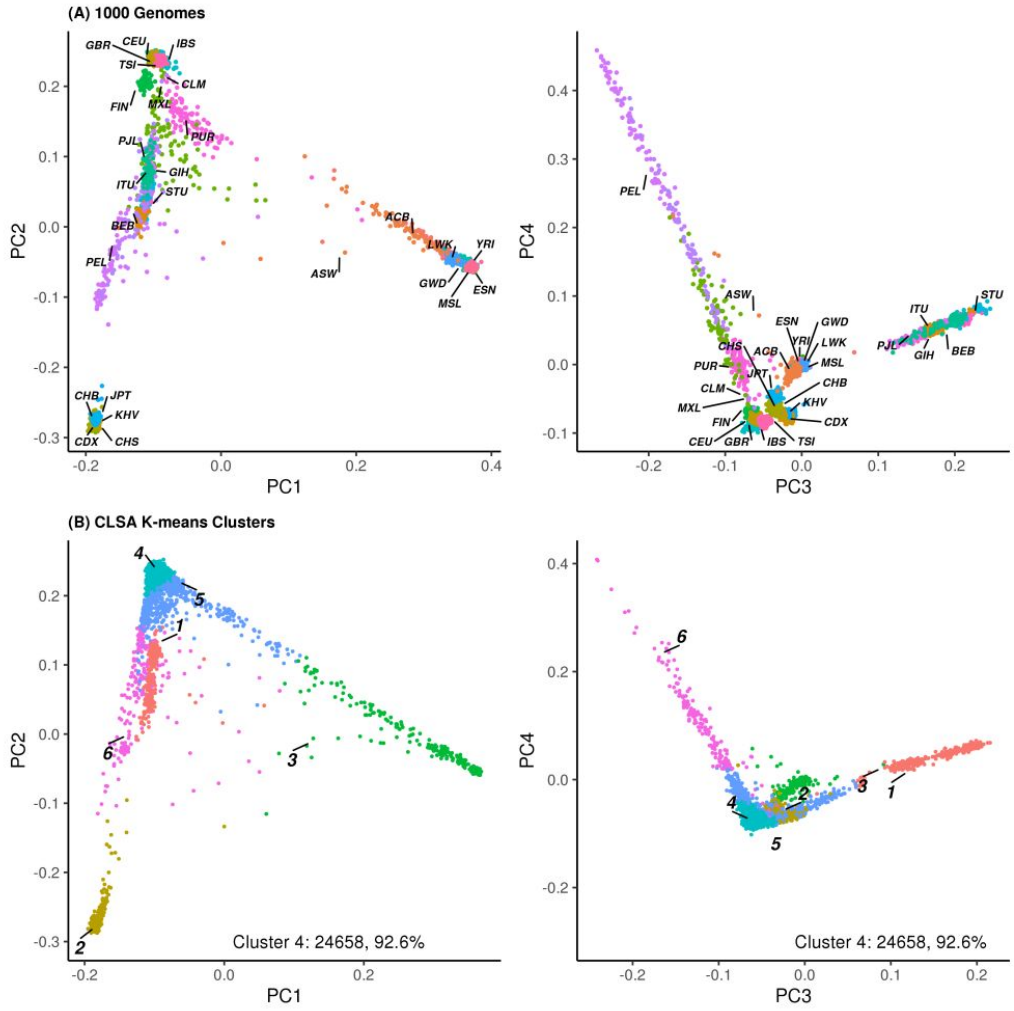


Figure 2: Determining the CLSA European ancestry subset.

(A) Top 4 principal components from all 1000 Genomes populations labelled and coloured. Population code refers to https://www.internationalgenome.org/category/population/. (B) Top 4 principal components from CLSA color coded and labelled by cluster number.

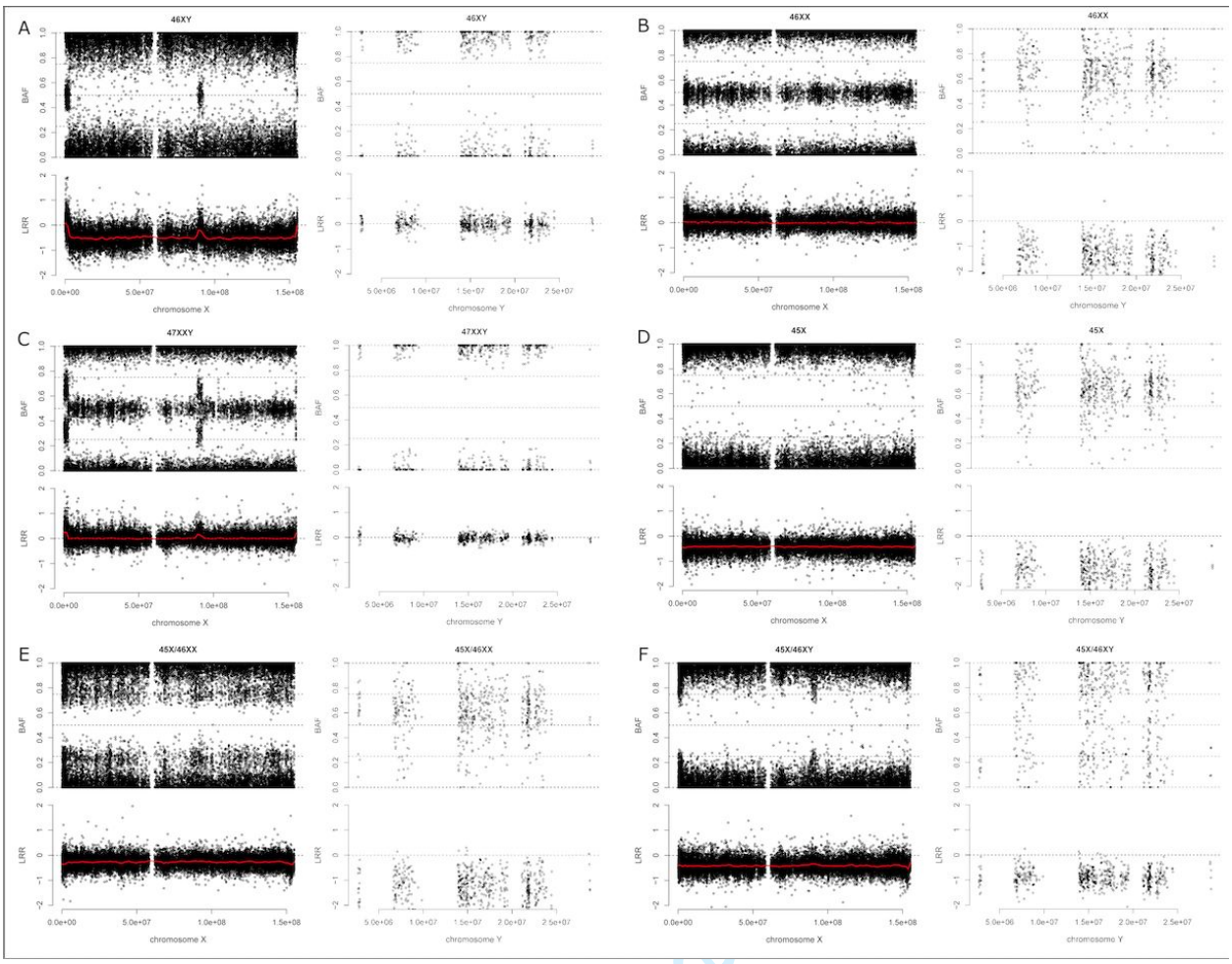
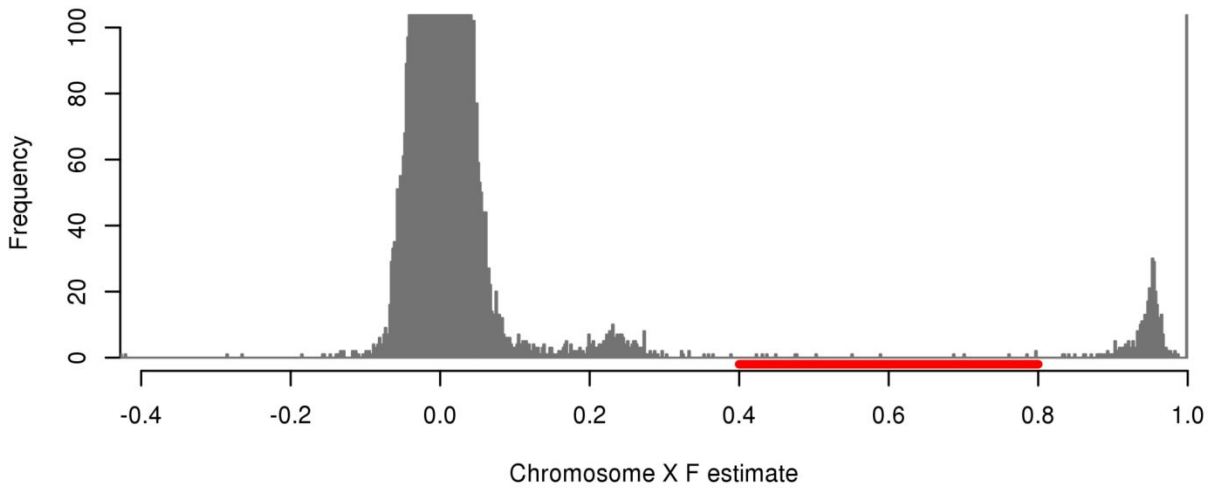
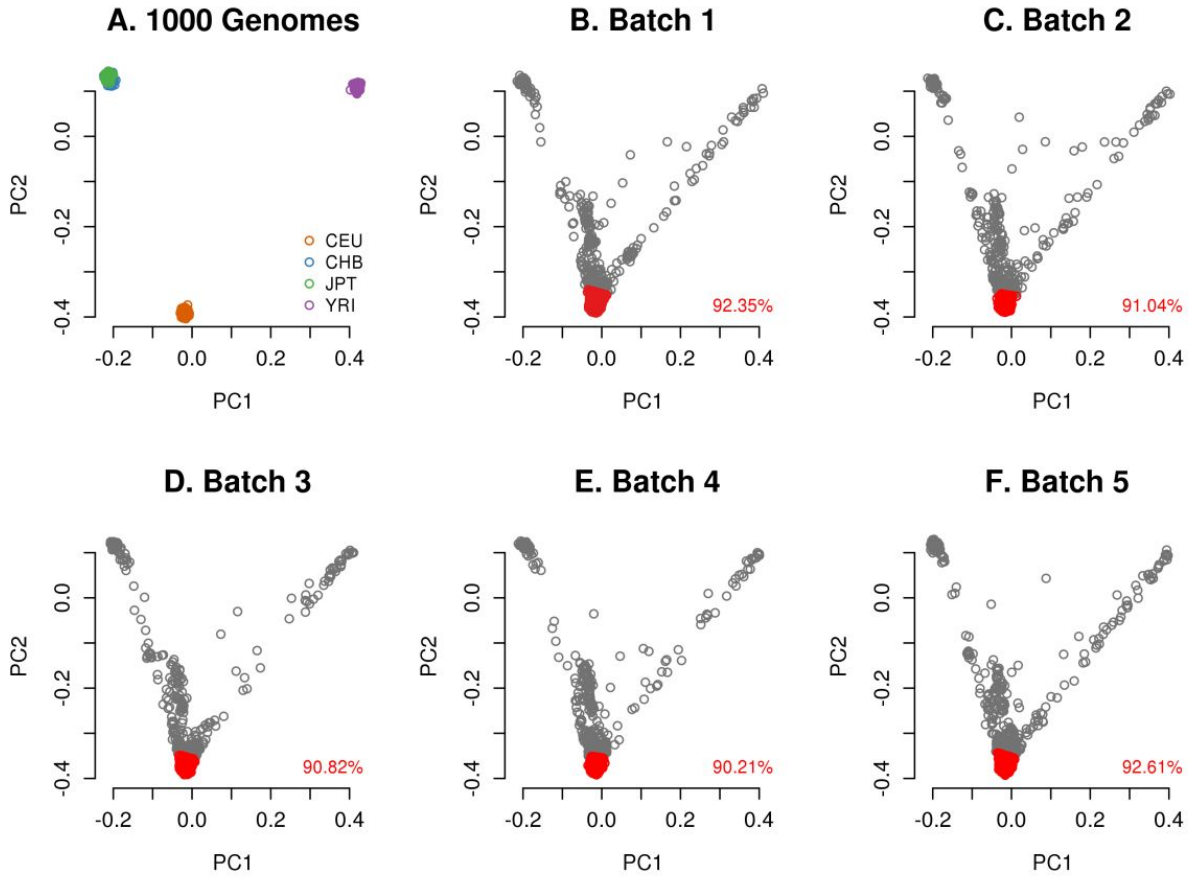


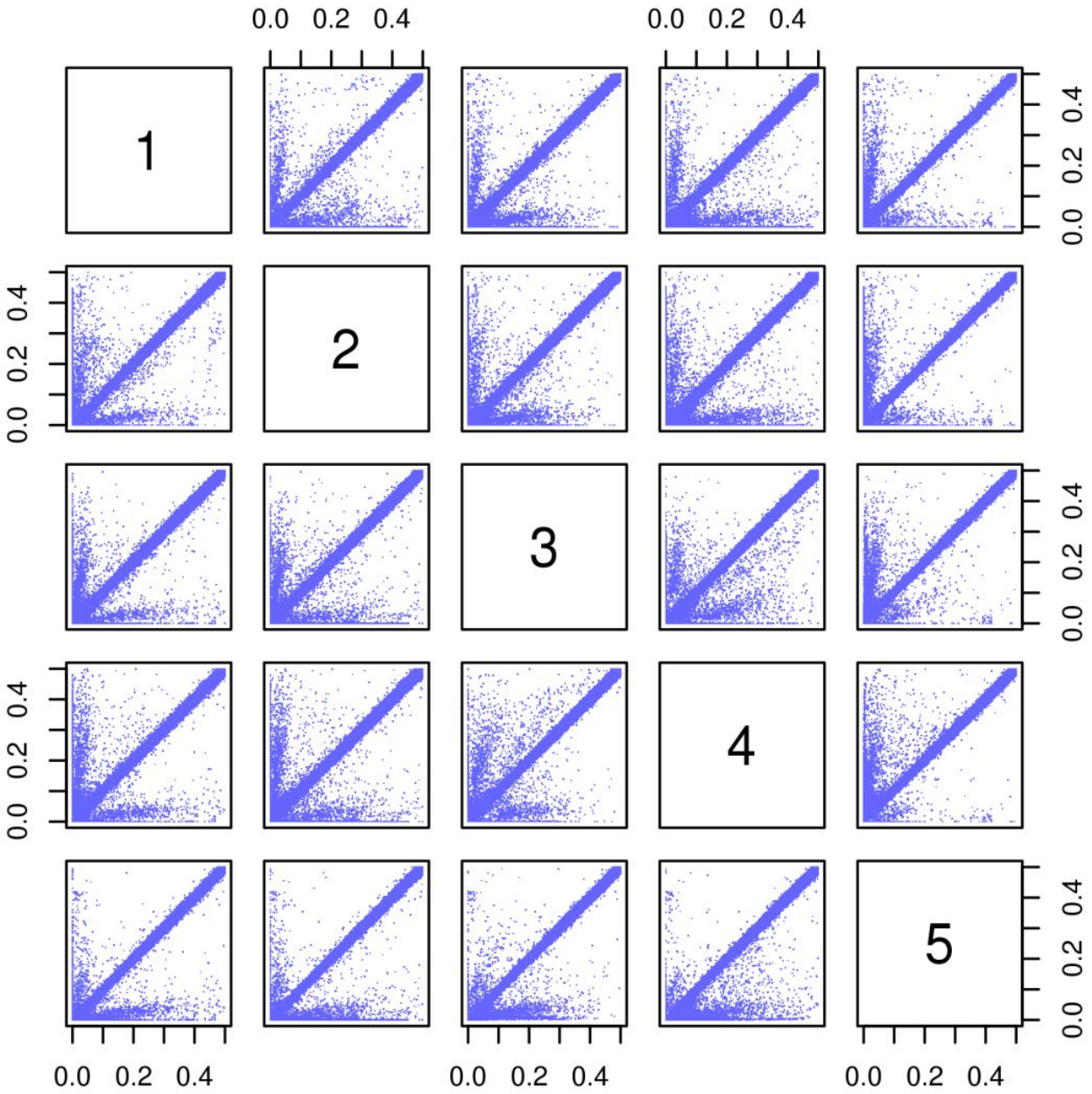
Figure 3: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X and Y are shown for sex chromosome abnormalities.



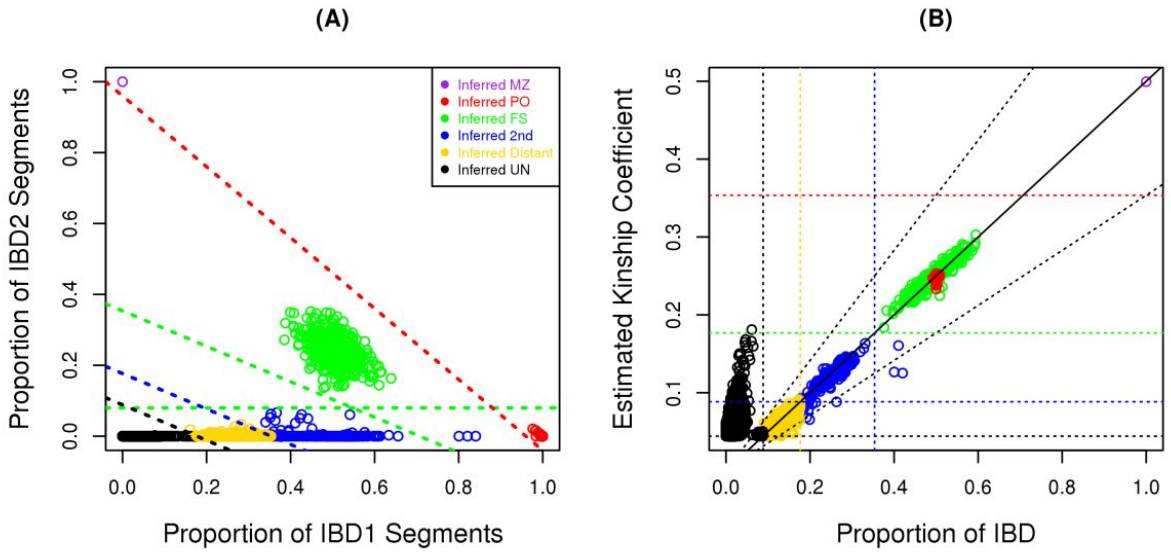
Supplementary Figure S1: Distribution of chromosome X F estimates for CLSA genotyped participants (y-axis truncated). Individuals with chromosome X F estimates within the range of 0.4 to 0.8 (red) are considered to have undefined chromosomal sex.



Supplementary Figure S2: 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).

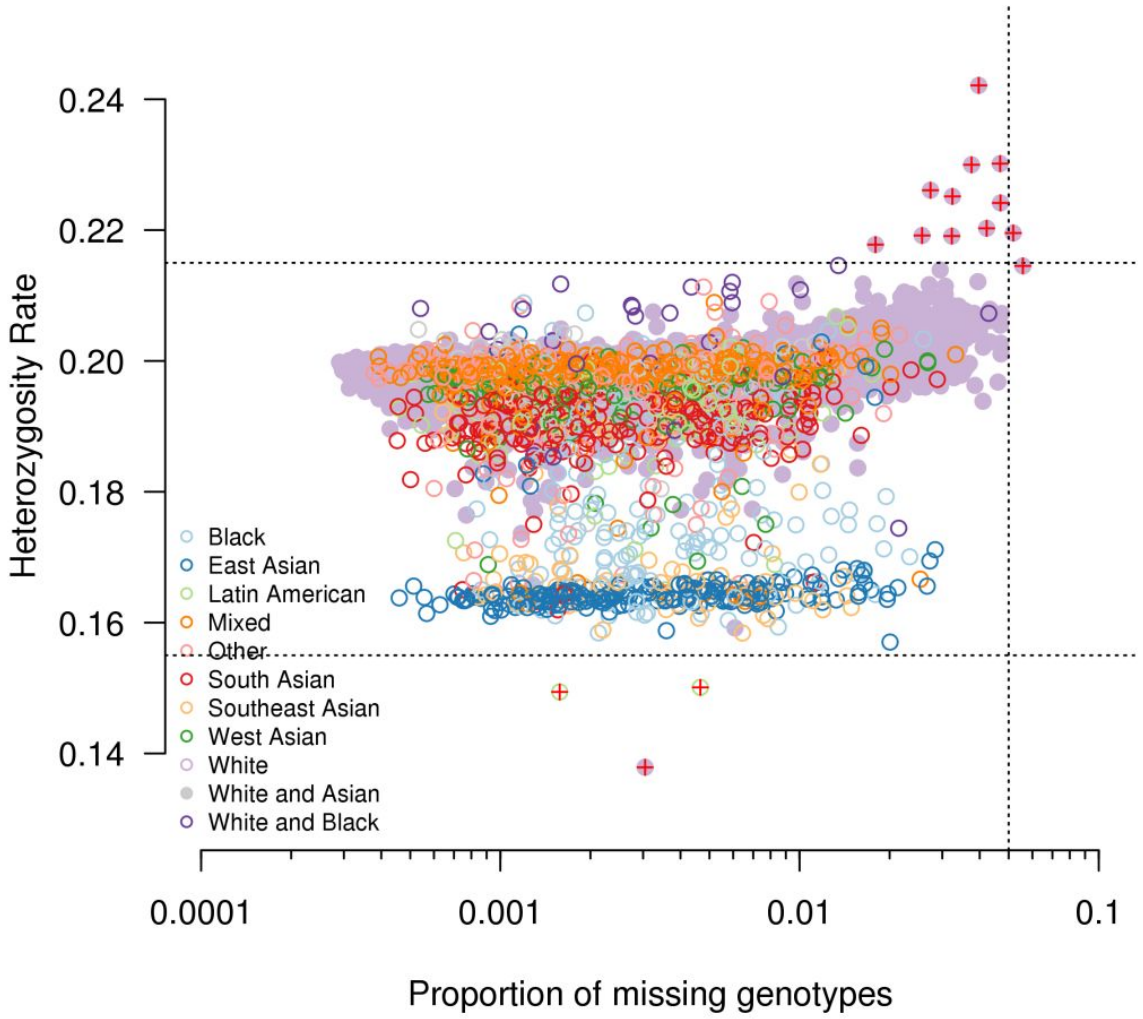


Supplementary Figure S3: Pairwise plot of allele frequency of SNPs from genotype batch 1 to 5.

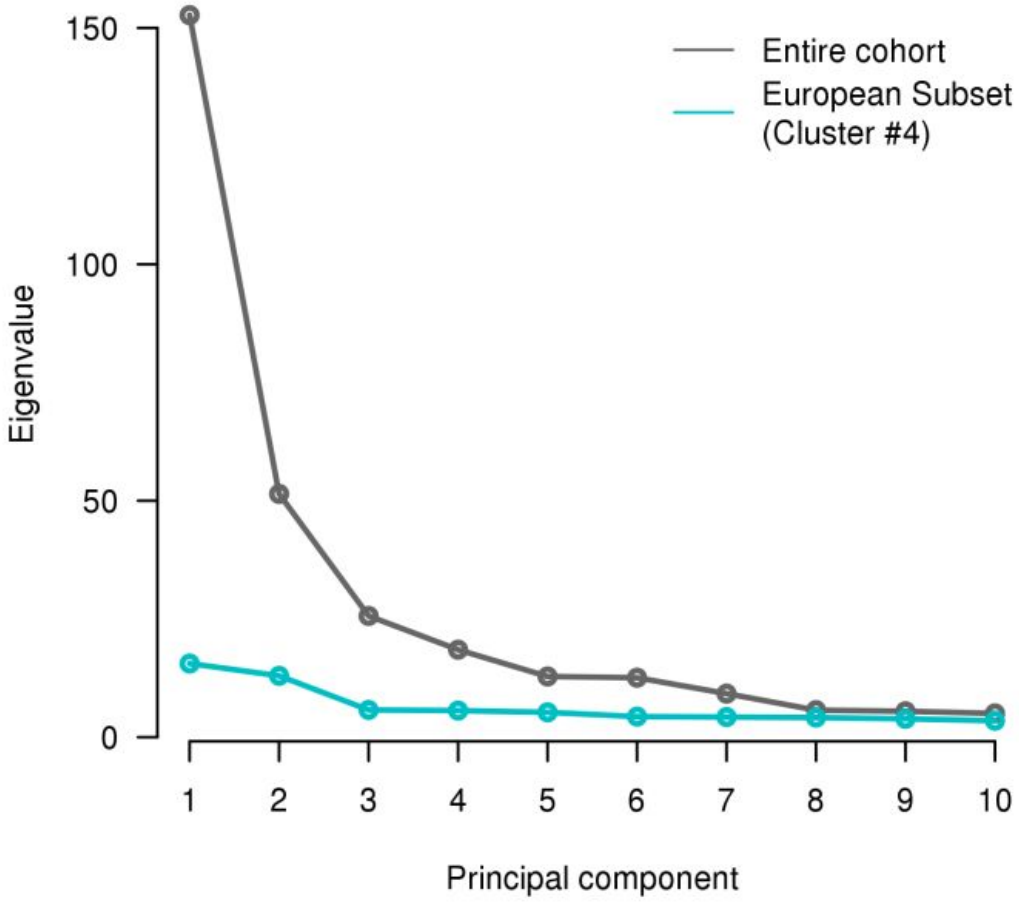


Supplementary Figure S4: Inference of familial relatedness using KING.

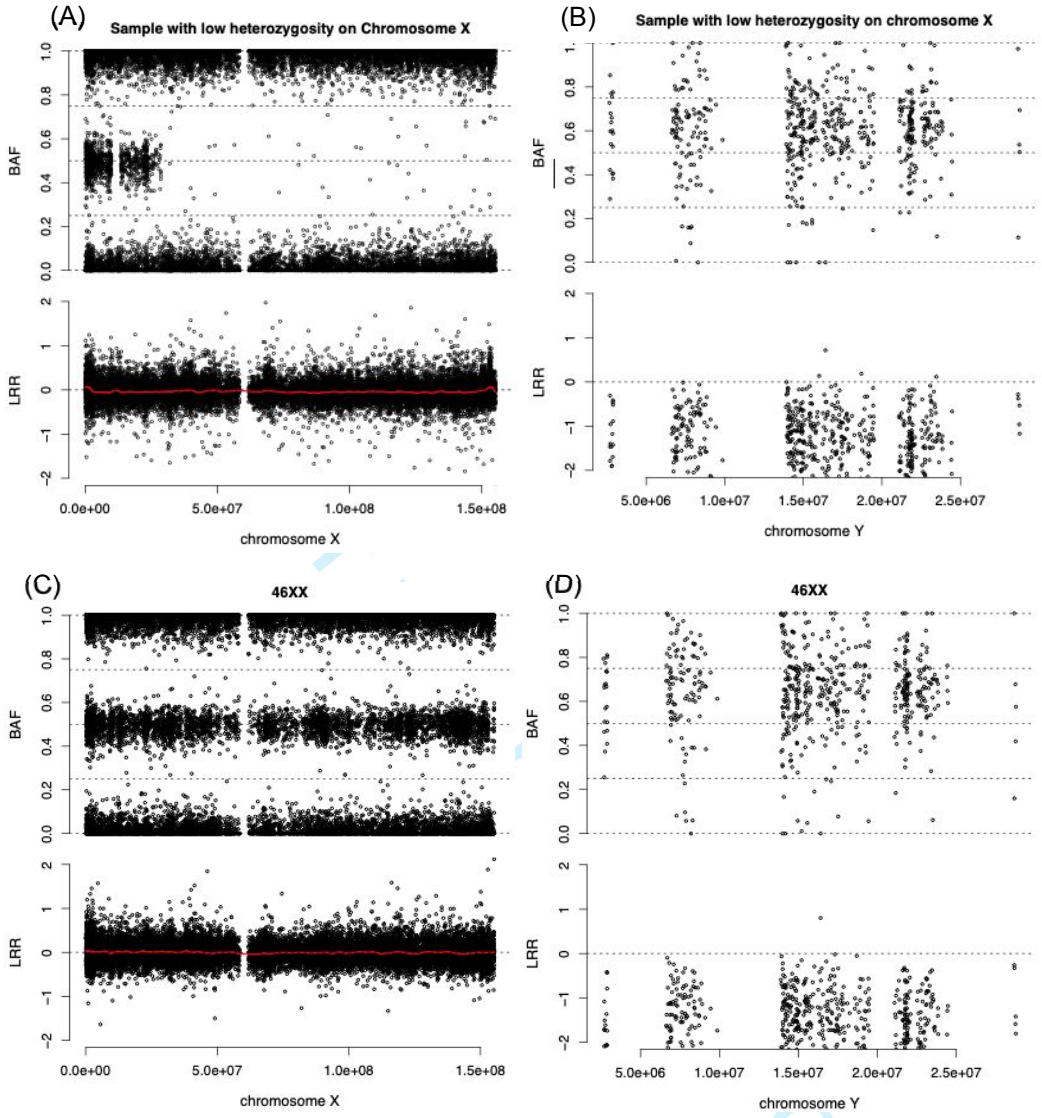
(A) Inference using IBD segments. (B) Inference using proportion IBD and kinship coefficient. Relationships in legend are abbreviated as: MZ=Monozygotic twin, PO=Parent/offspring, FS=Full sibling, 2nd=Second-degree relative, 3rd=Third-degree relative, Distant=Greater than 3rd degree relative, UN=Unrelated.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



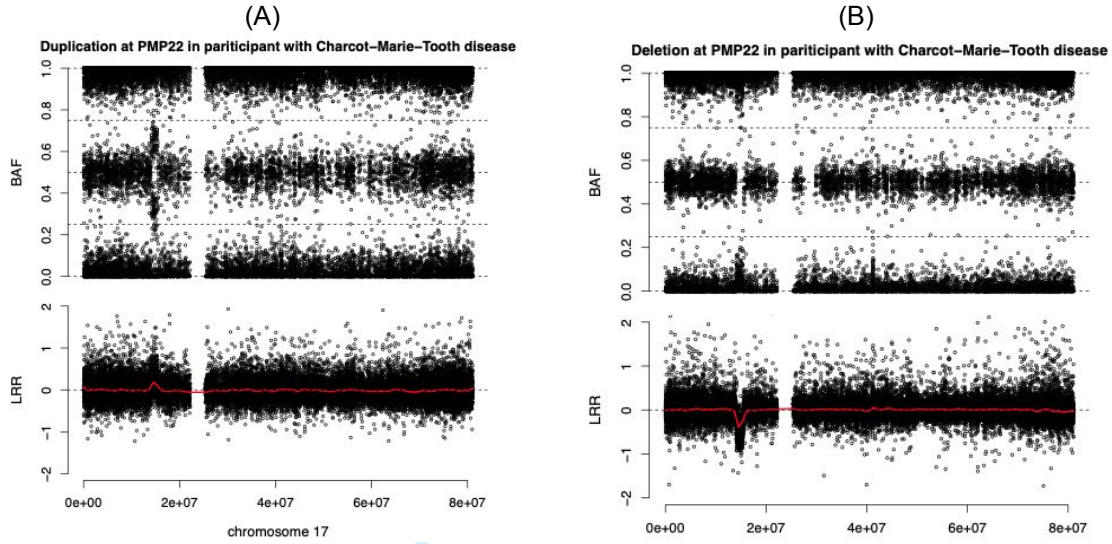
Supplementary Figure S5: Sample-wise heterozygosity versus genotype missingness. Points are color coded according to self-reported ancestry category. Outliers are marked with a red plus sign.



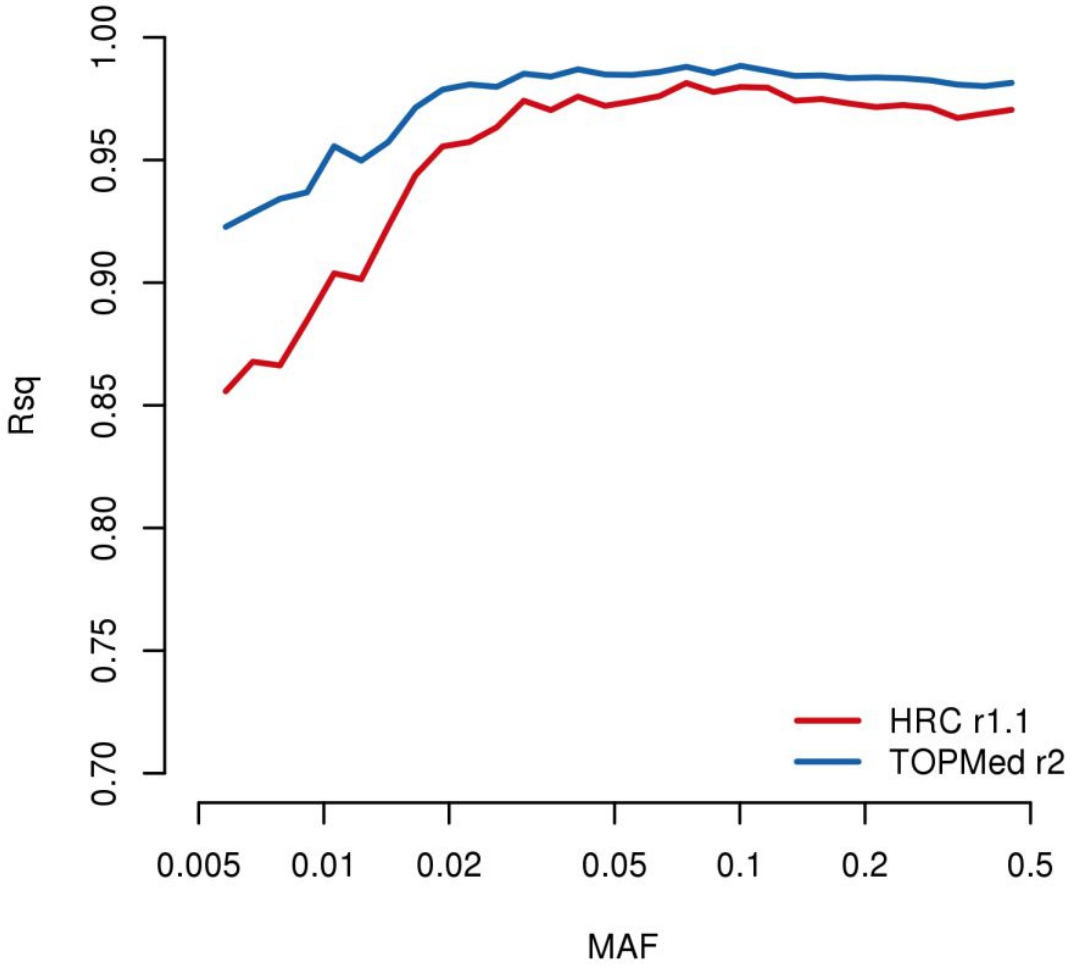
Supplementary Figure S6: Eigenvalues for PCA analysis of the entire cohort (grey) and the European ancestry subset (cluster 4, Robin egg blue), demonstrating a reduction in genetic variance within the European ancestry subset.



Supplementary Figure S7: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X (A) and Y (B) are shown for sample with low heterozygosity on chromosome X compared to sample with 46,XX (C-D).



Supplementary Figure S8: BAF (TOP) and log2 ratio (BOTTOM) of chromosome 17 are shown for sample with duplication (A) or deletion (B) at *PMP22* locus.



Supplementary Figure S9: Imputation quality of the CLSA cohort using the TOPMed versus Haplotype Reference Consortium (HRC) reference panel stratified by minor allele frequency (MAF) bins (data shown is from chromosome 22).

Supplementary Table S1: Sex chromosome determination of miscalled genotyped CLSA participant

Self reported sex	Affymetrix sex corrected by SVM	PLINK sex (raw F estimate <0.3 female raw F estimate >0.8 male)	discordance in section Sex chromosome composition	PLINK sex (adjusted F estimate <0.4 female adjusted F estimate >0.8 male)	sex determined by combined Affymetrix/PLINK/CNV approach	CLSA self-reported phenotype
male	female	female	1	female	female	
male	male	female	1	female	male	
female	male	male	1	male	male	
male	female	female	1	female	female	
female	female	unknown		female	female	
female	female	unknown	1	unknown	female	
female	female	unknown	1	unknown	female	
female	female	unknown	1	unknown	female	
male	female	female	1	female	female	
female	male	male	1	male	male	
female	male	male	1	male	male	
male	female	female	1	female	female	
female	female	male	1	male	female	Turner Syndrc
female	male	male	1	male	male	
female	female	unknown		female	female	
male	female	female	1	female	female	
female	male	male	1	male	male	
male	male	unknown	1	female	male	
female	male	male	1	male	male	
male	male	unknown	1	unknown	male	
male	male	unknown	1	unknown	male	
female	female	male	1	male	female	
female	male	male	1	male	male	
male	female	female	1	female	female	
female	female	male	1	male	female	
male	male	unknown		male	male	
female	female	unknown	1	unknown	female	
male	female	male		male	female	
female	female	unknown	1	unknown	female	
male	female	male		male	male	
female	female	male	1	unknown	female	
female	male	male	1	male	male	
female	male	unknown	1	male	female	Turner Syndrc
male	male	female	1	female	male	
female	female	male	1	male	female	
female	female	unknown		female	female	

male	male	unknown	1	unknown	male	
male	male	female	1	female	male	Klinefelter Syr
female	male	male	1	male	male	
male	male	unknown	1	female	male	
male	female	male		male	male	
female	female	male	1	male	female	Turner Syndrc
female	female	unknown		female	female	
female	female	unknown	1	unknown	female	
male	male	unknown		male	male	
female	female	unknown		female	female	
female	female	unknown	1	unknown	female	
female	female	unknown		female	female	
male	male	unknown	1	unknown	male	
male	female	male		male	male	
female	female	unknown	1	unknown	female	
male	female	female	1	female	female	
female	female	unknown		female	female	
female	male	male	1	male	male	
male	female	female	1	female	female	
male	male	female	1	female	male	Klinefelter Syr
female	female	unknown	1	unknown	female	
female	female	unknown		female	female	
male	female	female	1	female	female	
male	female	female	1	female	female	
female	male	male	1	male	male	
male	female	male		male	male	
male	male	unknown	1	unknown	male	

s

chromosomal abnormality from CNV profile	Raw F estimate	Adjusted F estimate
No abnormality	-0.00301	0.01745
Klinefelter Sync	0.1048	0.1002
No abnormality	1	1
No abnormality	0.01533	0.02791
Low heterozygc	0.3495	0.3644
45,X/46,XX mos	0.5095	0.4773
Low heterozygc	0.4637	0.4485
Low heterozygc	0.4352	0.4746
No abnormality	0.008892	0.03223
No abnormality	1	1
No abnormality	1	1
No abnormality	-0.03065	-0.02794
Turner Syndron	0.9507	0.9614
No abnormality	1	1
Low heterozygc	0.3043	0.2802
No abnormality	-0.002689	-0.01603
No abnormality	1	1
Klinefelter Sync	0.3124	0.3235
No abnormality	1	1
No abnormality	0.6854	0.6878
45,X/46,XY mos	0.5798	0.5893
Turner Syndron	0.9792	0.9875
No abnormality	1	1
No abnormality	0.04474	0.03819
Turner Syndron	0.9439	0.9545
45,X/46,XY mos	0.7965	0.8497
Low heterozygc	0.4524	0.4305
45,X/46,XY mos	0.9748	0.986
Low heterozygc	0.5457	0.5504
45,X/46,XY mos	0.9678	0.9797
Low heterozygc	0.8076	0.7966
No abnormality	1	1
45,X/46,XY mos	0.7827	0.8404
Klinefelter Sync	-0.0457	-0.04304
Turner Syndron	0.879	0.9028
Low heterozygc	0.3035	0.3339

45,X/46,XY mos	0.6732	0.7016
Klinefelter Sync	-0.03886	-0.02651
No abnormality	1	1
Klinefelter Sync	0.3093	0.2872
45,X/46,XY mos	1	0.9603
Turner Syndron	0.9273	0.9527
Low heterozygc	0.4004	0.3584
Low heterozygc	0.3838	0.4362
45,X/46,XY mos	0.7707	0.8358
Low heterozygc	0.3978	0.3886
Low heterozygc	0.7748	0.7842
No abnormality	0.3298	0.3527
45,X/46,XY mos	0.6658	0.7601
45,X/46,XY mos	1	0.9827
45,X/46,XX mos	0.4148	0.4228
No abnormality	-0.003668	-0.01753
Low heterozygc	0.3489	0.2975
No abnormality	1	1
No abnormality	-0.02345	-0.01932
Klinefelter Sync	-0.03581	-0.02924
45,X/46,XX mos	0.5139	0.5022
Low heterozygc	0.3461	0.3336
No abnormality	0.01936	-0.006017
No abnormality	-0.04206	-0.0411
No abnormality	1	1
45,X/46,XY mos	0.9756	0.9832
45,X/46,XY mos	0.72	0.7971

Supplementary Table S2: Self-reported ancestry and derived category from cultural and racial back

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

kground

Supplementary Table S3: Comparison of HLA types in positive controls with kown types in literatur

	NA10859 (134	NA10859 (134	NA10859 (1347-02)	NA12878 (146	NA12878 (146
HLA locus	Reference Ge	No of replicat	Accuracy(%)/Call rate(%)[d]	Reference Ge	No of replicat
A	03:01/01:01[b]	4	100/100	01:01/11:01[e]	587
B	07:02/15:01[b]	4	100/100	08:01/56:01[e]	587
C	06:02/07:02[b]	4	100/100	01:02/07:01[e]	587
DPA1	01:03/01:03[b]	4	100/100	01:03/02:01[e]	587
DPB1	04:02/04:02[b]	4	100/100	04:01/14:01[e]	587
DQA1	01:02/03:01[b]	4	100/100	01:01/05:01[e]	587
DQB1	03:02/06:02[b]	4	100/100	02:01/05:01[e]	587
DRB1	04:01/15:01[b]	4	100/100	01:01/03:01[e]	587
DRB3	NA[b,c]	4	-	01:01/01:01;0	587
DRB4	01[b]	4	100/100	01:01/01:01;0	587
DRB5	Na[b,c]	4	-	NA[e]	587

Note: a-Coriell ID (CEPH Family ID or NIST ID/RM Number for Personal Genome Project sample)

b: reference genotype source-IPD-IMGT/HLA Database

c: reference genotype data is not available

d: call rate is based on a posterior probability call threshold of 0.7

e: reference genotype source-PLoS Comput Biol. 2016 Oct; 12(10): e1005151. PMID: 27792722. A s

f: reference genotype source-DOI: 10.12688/f1000research.19630.1

g: reference genotype source-DOI: 10.12688/f1000research.19630.1 and https://www.pacb.com/v

h: reference genotype source-DOI: 10.12688/f1000research.19630.1 and Nature Communications

re

NA12878 (1463-02)	NA24385 (HG	NA24385 (HG	NA24385 (HG002)
Accuracy(%)/Call rate(%)	Reference Ge	No of replicat	Accuracy(%)/Call rate(%)
100/100	26:01/01:01[f]	75	100/100
100/100	38:01/35:08[f]	75	-
100/100	12:03/04:01[f]	75	100/100
100/100	01:03/01:03[f]	75	100/100
100/100	04:01/04:01[g]	75	100/100
100/100	03:01/01:01[h]	75	100/100
100/100	05:01/03:02[f]	75	100/100
100/100	04:02/10:01[f]	75	100/100
100/100	NA[f]	75	-
-	01:03[f]	75	100/100
-	NA[f]	75	-

set of possible alleles are reported in the reference. The HLA types we validated are shown in the ta

wp-content/uploads/Rowell-CSHLBioData-2018-Comprehensive-Variant-Detection-in-a-Human-Gen
doi: 10.1038/s41467-020-18564-9. The HLA types we validated are shown in the table.

ıble.

ıome-with-PacBio-High-Fidelity-Reads.pdf. The HLA types we validated are shown in the table.

BMJ Open

Cohort Profile: Genomic data for 26,622 individuals from the Canadian Longitudinal Study on Aging (CLSA)

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059021.R1
Article Type:	Cohort profile
Date Submitted by the Author:	07-Feb-2022
Complete List of Authors:	Forgetta, Vince; Jewish General Hospital, Centre for Clinical Epidemiology Li, Rui; McGill University, Darmond-Zwaig, Corinne; McGill University Belisle, Alexandre; McGill University Balion, Cynthia; McMaster University, Pathology and Molecular Medicine Roshandel, Delnaz; The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning Wolfson, Christina; McGill University Lettre, Guillaume; Université de Montréal; Montreal Heart Institute Pare, Guillaume ; McMaster University Paterson, Andrew; Hospital for Sick Children, Griffith, Lauren; McMaster University, Department of Health Research Methods, Evidence, and Impact Verschoor, Chris; McMaster University, Lathrop, Mark; McGill University, Department of Human Genetics Kirkland, Susan ; Dalhousie University, Raina, Parminder; McMaster University, Clinical Epidemiology and Biostatistics Richards, Brent ; McGill University, Ragoussis, Jiannis; McGill University, Department of Human Genetics; McGill Genome Centre
Primary Subject Heading:	Genetics and genomics
Secondary Subject Heading:	Genetics and genomics, Epidemiology, Public health, Qualitative research
Keywords:	GENETICS, EPIDEMIOLOGY, PUBLIC HEALTH, Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Glaucoma < OPHTHALMOLOGY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Cohort Profile: Genomic data for 26,622 individuals from the Canadian Longitudinal Study on Aging (CLSA)

Author List:

Vincenzo Forgetta[1†], Rui Li[2†], Corinne Darmond-Zwaig[2], Alexandre Belisle[2], Cynthia Balion[3], Delnaz Roshandel[4], Christina Wolfson[5], Guillaume Lettre[6], Guillaume Pare[3], Andrew D. Paterson[4,7,8], Lauren E. Griffith[9], Chris Verschoor[9], Mark Lathrop[2], Susan Kirkland[10], Parminder Raina[9‡], J. Brent Richards[1,5,11,12‡], and Jiannis Ragoussis[2,12,13‡]

1 Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montréal, QC, Canada,

2 McGill University Genome Centre, Department of Human Genetics, McGill University, Montréal, QC, Canada,

3 Hamilton Regional Laboratory Medicine Program, McMaster University, St. Joseph's Hospital St. Luke's Wing, Hamilton, ON, Canada,

4 Genetics & Genomic Biology, The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, ON, Canada,

5 Department of Medicine, & of Epidemiology and Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada,

6 Montréal Heart Institute and Université de Montréal, Montréal, QC, Canada,

7 Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada,

8 https://orcid.org/0000-0002-9169-118X

9 Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada,

10 Department of Community Health and Epidemiology, Division of Geriatric Medicine, Dalhousie University, Halifax, Nova Scotia, Canada,

11 Department of Twin Research and Genetic Epidemiology, King's College London, London, UK,

12 Department of Human Genetics, McGill University, Montréal, QC, Canada,

13 Department of Bioengineering, McGill University, Montréal, QC, Canada,

* Corresponding author. McGill Genome Centre, 740 Avenue Dr. Penfield, Montreal, Québec, Canada H3A 0G1. Email: ioannis.ragoussis@mcgill.ca.

† Joint first authors.

‡ Joint senior authors.

Keyword: CLSA, genome-wide genotyping, aging, HLA

Word count: 4,435

Abstract

Purpose: The Canadian Longitudinal Study on Aging (CLSA) Comprehensive cohort was established to provide unique opportunities to study the genetic and environmental contributions to human disease as well as aging process. The aim of this report is to describe the genomic data included in CLSA.

Participants: A total of 26,622 individuals from CLSA comprehensive cohort of men and women aged 45 to 85 recruited between 2010 and 2015 have undergone genome-wide genotyping of DNA samples collected from blood. Comprehensive quality control metrics were measured for genetic markers and samples respectively. The genotypes were imputed to the TOPMed reference panel. Sex chromosome abnormalities were identified by copy number profiling. Classical HLA genes haplotypes were imputed at two-field (four-digit).

Findings to date: Of the 26,622 genotyped participants, 24,655 (92.6%) were identified as having European ancestry. This genomic data is linked to physical, lifestyle, medical, economic, environmental, and psychosocial factors collected longitudinally in CLSA. The combined

analysis including CLSA genomic data uncovered over 100 novel loci associated with key parameters to define glaucoma. The CLSA genomic dataset validated the contribution of a polygenic risk score to screen individuals with high fracture risk. It is also a valuable resource to directly identify common genetic variations associated with conditions related to complex traits. Taking advantage of the comprehensive interview and physical information collected in CLSA, this genomic dataset has been linked to psychosocial factors to investigate both the independent and interactive effects on cardiovascular disease.

Future plans: The CLSA overall is ongoing. Follow-up data will continue to be collected from participants in the current genomic subcohort including the DNA methylation and metabolomic data. Ongoing studies focus on elucidating the role of genetic factors in cognitive decline and cardiovascular diseases. This genomic data resource is available upon request through the CLSA data access application process.

Strengths and limitations of this study

- The genomic data in Canadian Longitudinal Study on Aging (CLSA) Comprehensive cohort provides whole-genome genotyping data on 794,409 markers and whole-genome imputed data on approximately 308 million genetic variants.
- The UK Biobank array used for genotyping is enriched with markers associated with multiple phenotypes including the comprehensive pharmacogenomic and inflammation markers which may be of particular interest since DNA methylation, metabolomic and proteomic data are being generated by CLSA.
- The CLSA cohort continues to follow up the participants on a wide spectrum of qualitative and quantitative variables; it will facilitate research on the effect of interplay between genetics and environmental factors on age-related diseases.
- Potential limitations may include the relatively lower genotyping coverage in participants with non-European ancestry, which can be substantially improved by using imputation

reference panel with high diversity and inadequate power to discover very rare predisposition variants.

Introduction

The global life expectancy increased dramatically through the past two hundred years. In such times, the make-up of Canadian population has changed unprecedentedly. From 1977 to 2017, the senior population, i.e., people aged 65 years and older, grew from 2 million to 6.2 million, which equaled to nearly 17% of its population size. This number is still rapidly rising. It is anticipated that by 2036 there will be 10.2 million senior people in Canada. Of every 4 Canadians, there will be one senior person.

Along with the expanded human life expectancy, the prevalence of age-related diseases is strikingly increasing. Aged people experience progressive decline in functional integrity and homeostasis. This process is accompanied by increased risk of neurodegeneration, cardiovascular disease and cancer among many other diseases, which have become the most common causes of decreased life quality and late-life mortality. It adds substantial burden to individual and social health care system inadvertently. Age-related diseases have a highly complex nature. Both the genetic and environmental factors play an important role as well as the interaction between them [1 2]. Therefore, understanding of the underlying mechanisms of aging is required for sustaining longer lives with reduced loss of healthy years.

Studies on short-lived model organisms provided insights on several key genetical regulators in hallmark aging pathways, however, the identification of biomarkers of age and age-related disease in human is more complicated [3]. Over the past decades, genetic epidemiology methods emerged to be a powerful tool. The genome-wide association studies (GWASs) have uncovered tens of genes and genetic variations that play a role in the variability of aging outcomes among people [4]. However, the genetic effects are usually relatively moderate and can be altered by lifestyle and other environmental determinants [2 5]. More work is needed to fully deconvolute the

interplay between genetics and extrinsic influences. This effort will be benefited by larger sample size and linked information on proteomics and epigenetics.

Cohort description

The Canadian Longitudinal Study on Aging (CLSA) is a national long-term study that recruited 51,338 men and women, aged 45-85 years at enrolment between 2010 and 2015 for baseline data collection [6]. It presents a unique opportunity to study the 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. It is composed of two complementary cohorts: the Tracking cohort of 21,241 participants who were interviewed by telephone and the Comprehensive cohort of 30,097 participants who were interviewed in person and provided blood and urine samples. The participants in the Comprehensive cohort were randomly selected from within 25-50 km of 11 data collection sites in seven provinces. A total of 27,170 (90.3%) Comprehensive cohort participants provided blood samples at baseline. The Comprehensive Cohort samples have been used to produce whole genome genotyping data. The data were collected to understand, individually and in combination, the impact of genetic variation in both maintaining health and in the development of disease and disability as people age. In this release of the CLSA genomic data, 26,622 participants have been genotyped using the Affymetrix UK Biobank Axiom array [7]. Qualified researchers from any country can access these genomic and phenotypic data via a formal data and sample access procedure described on the CLSA website (https://www.clsa-elcv.ca/data-access).

Patient and public involvement

None.

Data collected:

Sample storage and DNA extraction

The CLSA protocol was reviewed and approved by 13 research ethics boards across Canada. All participants provided written informed consent [8]. The biological samples were collected at the Data Collection Sites and de-identified. Whole blood buffy coats were isolated from peripheral blood drawn and the plasma layer was removed. Samples were immediately moved to -80°C storage, and transferred to liquid N_2 storage at the CLSA Biorepository and Bioanalysis Centre 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, Baesweiler, Germany). All extracted samples were quantified using PicoGreen Reagent Kit (Life Technologies, catalog # P7589). A minimum DNA 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.

Genotyping and calling

Each plate genotyped contained 92 CLSA DNA samples and 4 controls, one male control as the Affymetrix Reference Genomic DNA 103 (Catalog# 900421) or Personal Genome Project sample huAA53E0 (Coriell Cell Repositories, catalog # NA24385), two female controls as the CEPH control 1463-02 (Coriell Cell Repositories, catalog # NA12878) or the CEPH control 1347-2 (Coriell Cell Repositories, catalog # NA10859), and a deionized water negative control. The Affymetrix protocol (Axiom 2.0 Assay Automated workflow on Affymetrix NIMBUS) was followed. Samples were hybridized to UK Biobank arrays (ThermoFisher Catalog #902502), the same array that was used to genotype ~450,000 individuals in the UK Biobank [9].

Axiom Array plates were processed on the Affymetrix GeneTitan Multi-Channel Instrument. For first pass quality control (QC), batches of 8 plates were analyzed using the Sample QC workflow

of the AxiomTM Analysis Suite 2.0 software where a subset of 20,000 reliable probes were used to determine the resolution of the AT and GC signal contrast (Dish QC) and sample QC. The reliable probes are autosomal, previously wet-lab tested by the provider, working probe sets with two array features per probe set.

Genotyping quality control and removal of duplicate genotyped participants

Genotyping was undertaken in separate batches of approximately 5,000 samples each using AxiomTM Analysis Suite 2.0, similar to UK Biobank genotyping QC documentation [7]. Genotype calling resulted in 27,010 successfully genotyped DNA samples. An inclusion list containing 794,409 genetic variants was used [9], as well as the following QC parameters for selecting samples passing to further analysis: Dish QC ≥ 0.82 on sample level, and average QC call rate of passing samples on a plate (plate QC call rate) ≥ 95%, percentage of passing samples ≥ 70%, and average call rate for passing samples ≥ 95% on plate-level. Duplicate genotyped participants were detected by KING version 2.1.3 [10] and the sample with higher genotype missingness was removed. This resulted in 26,622 successfully genotyped participants.

Sex chromosome composition

Distribution of F estimates on the X chromosome showed a gap between 0.4 and 0.8 (Supplementary Figure S1). Using this threshold, we obtained X chromosome number using PLINK version 1.90b4.4 [11 12]. F estimates for the 48 individuals with sex discrepancies between self-reported sex and X chromosome composition (Table 1) are listed in Supplementary Table S1. All subsequent analyses in this paper will use X chromosome number and number of nonmissing Y chromosome genotypes to define sex.

Genetic marker-based quality control

This 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.

The above tests require a population with relatively homogenous ancestry. Given this, we determined the largest subset of ancestrally homogeneous participants via K-means clustering of projected principal components from 414 individuals across 4 populations (Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT) and Yoruba in Ibadan, Nigeria (YRI)) from 1000 Genomes Phase 3[13]. The largest cluster across all genotype batches overlapped the CEU population and included 24,361 individuals, or 92% of the entire genotyped cohort (N=26,622) (Supplementary Figure S2).

We then set 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 0.001 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. Single nucleotide polymorphisms (SNPs) that failed the tested QC parameters are flagged within the marker quality table provided with the data release. We thus invite researchers to filter markers based on these properties or devise their own quality control metrics that satisfy their research requirements.

Discordant genotype frequency between batches

To detect deviation in genotype frequency of markers between batches, we used a Fisher's exact test on the 2x3 table of genotype counts (or 2x2 table for haploid markers). The vast majority of markers did not exhibit significant deviation in genotype frequency (779,656, 98.1% of total).

Departure from Hardy-Weinberg equilibrium

We conducted the test for departure from HWE using the exact test [14]. There were 7,790 markers with an HWE p-value < 3.15×10^{-10}.

Discordance across control replicates

There were 3 positive control samples on each genotyping plate: a male control (Affymetrix CTL1 103 or Personal Genome Project participant huAA53E0), and one of two female controls (CEPH 1463-02 or CEPH 1347-02) in duplicate. For each marker and control sample we computed a discordance metric (d) defined as 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. There were 27,937 markers with control replicate discordance greater than 0.05 (i.e. concordance < 0.95).

Sex genotype frequency discordance

To detect deviation in genotype frequency of markers between sexes, we used Fisher's exact test on the 2x3 table of genotype counts for autosomal SNPs (or 2x2 table of allele counts for the sex-specific regions of the X chromosome). There were 248 markers with discordant genotype counts or allele counts between sexes with p-value < 3.15×10^{-10}, in which 192 markers were on sex-specific region of the X chromosome.

Summary of results from marker-based tests

There were 37,706 SNPs that were flagged by one or more of the 4 tests. They are labeled in the marker quality control file accompanying this data release. The effect of this quality analysis is depicted by comparing Supplementary Figure S3 with Figure 1 where there is clear improvement in the concordance in minor allele frequency (MAF) between batches after removal of these markers. We recommend removing these markers but have maintained these markers in the dataset so that researchers have access to all data. In addition, 15,616

insertions/deletions and 95,363 low-frequency SNPs with MAF < 0.005 were flagged as they may bias subsequent sample-based quality control.

Sample-based quality control

This sample-based quality control was intended to identify samples of low-quality, related individuals, and provide a genetic-based description of ancestry. We thus encourage researchers using this information included in the data release to filter samples or devise their own sample quality control metrics that satisfy their research requirements.

We selected the SNPs that passed all 4 tests from marker-based quality control with MAF > 0.01 and marker-wise missingness < 0.01 resulting in a total of 573,386 markers. PLINK was used to prune these markers to a subset of 161,536 independent markers in approximate linkage equilibrium. They were used for the following sample-wise assessments. The pruning was done on window size of 5000 kb with pairwise r^2 threshold as 0.1 and the number of variants to shift the window as 5.

Familial relatedness

Familial relationships among CLSA participants were not recorded in the questionnaires or interviews. However, this information is essential for some epidemiological and genomic analyses. Using the KING software [10] we computed all pairwise kinship coefficients and noted all pairs with inferred relatedness of 3rd degree or closer using autosomal SNPs (Table 2, Supplementary Figure S4). Individuals with an inferred relationship of 3rd degree or closer are labeled in the database.

Detection of outliers in heterozygosity and missing rates

Since extreme values in sample-wise heterozygosity and missingness may suggest low quality genotyping or cross-contamination of biological samples, we detected outliers by using PLINK (Supplementary Figure S5). As expected, because the allele frequencies differ between populations, we observed that heterozygosity was dependent on self-reported background.

Population structure

Population structure was computed by principal component analysis (PCA) [15] to complement self-reported ancestry and control for population stratification in GWAS [16 17]. The top 20 principal components were computed using a high-quality subset of unrelated individuals by removing individuals classified as outliers in heterozygosity and missingness, and any individual with a relation of 3rd degree or less.

Selection of 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). We combined self-reported ancestry with genomic information and PCA analysis to identify a subset of self-reported European individuals with relatively homogenous 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 into 6 clusters. Visualization of these clusters alongside those from 1000 Genomes reveals a clear overlap of the largest cluster (cluster 4, N=24,655) with populations of European ancestry in 1000 Genomes (Figure 2). Moreover, this largest cluster contains the vast majority of individuals in CLSA that self-report European ancestry (Table 3, Supplementary Table S2). The European ancestry subset has markedly reduced variance in the top principal components as compared to the entire CLSA cohort (Supplementary Figure S6). 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.

Detection of copy number abnormalities associated with disease

Sex chromosome abnormalities

The sex chromosome composition was called by both Affymetrix Axiom™ Analysis Suite 2.0 and PLINK. Affymetrix uses the ratio of mean signal values of non-polymorphic probes separately on the X and Y chromosomes to calculate sex. PLINK determines sex by using only X chromosome inbreeding coefficient (F estimates). When a subject has sex chromosome abnormalities such as Turner syndrome (45, X), Affymetrix will call them female but PLINK will call them male. Similarly, when a subject has Klinefelter Syndrome (47, XXY), Affymetrix will call the subject male but PLINK will call them female. We use this discordance information combined with copy number profiling to identify sex chromosome abnormalities in CLSA participants.

To correct the miscalling of males by stringent Affymetrix default threshold, the intensity data of chromosome X and Y markers from all UK Biobank samples were used as a training data set to generate a Support Vector Machine (SVM) model. This SVM model was applied to CLSA samples to recall the vast majority of miscalled samples (331 out of 359). However, the SVM approach as aforementioned could not be applied to PLINK sex calling since the sex calling in UK Biobank data was already corrected. Alternatively, an empirical threshold was used to recall most (140 out of 175) of the samples miscalled by PLINK through setting X chromosome F estimate < 0.3 as female and > 0.8 as male. We used a relatively more stringent threshold of F estimate because high F estimates may indicate mosaic chromosomal abnormalities such as mosaic deletion. Finally, we used Axiom CNV Summary Tool to calculate log2 ratio and B allele frequency (BAF, which is in fact the within person ratio of B/B+A intensity at each SNP) for both X and Y chromosomes from the genotyping data. The log2 ratio and BAF were used to identify sex chromosome abnormalities compared to males and females with 46,XY and 46,XX, respectively (Figure 3 (A-B)).

As a result, we detected 63 participants with discordance between self-reported sex and Affymetrix and/or PLINK sex calling (Supplementary Table S2), then we examined their CNV to identify them as one of four scenarios, sex chromosome aneuploidy (11 subjects), mosaic sex

chromosome aneuploidy (15 subjects), low heterozygosity on the X chromosome (14 subjects), discordance between X chromosome number and self-reported sex without sex chromosome aneuploidy (23 subjects). Briefly, we identified all 5 participants with self-reported sex chromosome abnormalities including 1 mosaic Turner syndrome patient (45,X/46,XY) (scenarios 1 and 2). We identified all 48 participants with sex discordance as in abovementioned sex check. For the 23 participants who had discordance with both Affymetrix and PLINK calling, CNV analysis confirmed the sex chromosome composition (scenario 4). In addition, for participants with no self-reported sex, Affymetrix/PLINK calling and CNV analysis are concordant to call sex. Besides the validated self-reported sex chromosomal abnormalities, we identified 4 participants with Klinefelter syndrome (47,XXY) and 3 with Turner Syndrome (45,X) (scenario 1) (Figure 3 (C-D)). In total, we found 3 participants with 45,X/46,XX mosaicism, and 11 participants with 45,X/46,XY mosaicism including 1 with self-reported Turner syndrome (45,X/46,XY) (Figure 3 (E-F)). Additionally, individuals with low heterozygosity on the chromosome X could be a result of inbreeding (Supplementary Figure S7).

Charcot-Marie-Tooth Disease

Charcot-Marie-Tooth disease (CMT) is one of the most common inherited neurological disorders. It is mostly caused by duplication at 17p12 where *PMP22* is located (CMT1A and CMT1E; OMIM: # 118220; # 118300). In this release of CLSA genomic data, there are 9 CLSA participants who self-reported as having CMT. We examined their CNVs and found that 4 participants have duplication at *PMP22* (Supplementary Figure S8), and 1 participant has deletion at *PMP22* (Supplementary Figure S8). The other 4 subjects did not have CNVs detected at *PMP22*.

HLA type imputation

We used the HLA*IMP:02 method [18] and a multi-population reference panel [18] (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. Calling accuracy was 100% across the loci (Supplementary Table S3). The imputation accuracy of genotyped CLSA participants was estimated by using the replicated samples. The validation rate is 100% for all the replicates.

Imputation to the TOPMed reference panel

Genotype imputation is a computational method to predict marker genotypes that are not directly genotyped by an assay, such as genotyping array, or to impute markers that are missing in certain individuals. The imputation process uses a reference panel of sequenced individuals to predict genotypes in a study sample for which only a subset of these genetic markers has been genotyped [19]. As input to the imputation process, we used the 26,622 CLSA participants that passed quality control, and the set of 653,729 markers that passed all marker QC tests, with SNP-wise missingness < 0.05, MAF > 0.0001 and have alleles that match the human genome GRCh37 reference sequence.

Phasing and imputation were conducted using the TOPMed reference panel [20] at the University of Michigan Imputation Service [21]. 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 [22] and Minimac [19], 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 3 control samples. The two batches where subsequently merged into a single dataset.

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 [23]. For each imputation data set, information measures for all SNPs on chromosome 22 were stratified into 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 (Supplementary Figure S9). The relatively better imputation performance may also be empowered by the higher sequencing depth and joint calling method that were used to generate the TOPMed reference panel.

Findings to date

This data resource has been used in four completed and several ongoing studies. Glaucoma is the second leading cause of irreversible blindness in the world [24]. The GWAS combining data from UK Biobank, CLSA and the International Glaucoma Genetic Consortium identified more than 100 novel loci for vertical cup-to-disc ratio and vertical disc diameter [25]. They are highly heritable optic disc morphology traits related to glaucoma risk. In a study to investigate the contribution of polygenic risk score (PRS) to screening for fracture risk [26], the CLSA genomic data were linked to the participants' physical examinations. It was the largest cohort included in this combined analysis of fracture risk, which enabled the researchers to understand the performance of PRS particularly in older individuals. It was found that the genetic pre-screening could reduce the number of further assessments to identify individuals at high risk of osteoporotic fractures. In another study on cardiovascular disease [27], the investigators evaluated the independent effects and interactions of multiscale risk factors by taking advantage of combined genomic and psychosocial information collected in CLSA cohort. In addition, the CLSA dataset provides opportunities to study other conditions related to complex diseases. It was employed by a large scale GWAS on sleep apnoea which was associated with cardiovascular disease and glaucoma. The authors revealed robust novel associations between

30 genes and this condition, and substantial molecular overlap with other complex traits [28]. For further publications please consult https://www.clsa-elcv.ca/stay-informed/publications.

Strengths and limitations

The CLSA genomic data are a unique resource nested in a large-scale, longitudinal study profiling the aging population in Canada. The genotyping array is enriched with known markers associated with multiple phenotypes. However, the UK Biobank array may have relatively lower coverage in participants with non-European ancestry [29], which can be improved by using imputation reference panels with high genetic diversity [30]. It may be difficult to identify very rare variants by using this genotyping data since the current imputation method cannot confidently predict variants with frequency under certain threshold. In spite of these limitations, CLSA cohort includes deep and extensive phenotyping and planned linkage to health administrative databases. For example, recently the metabolomic data comprising 1,314 biochemicals became available in approximately 9,500 blood samples collected from CLSA participants, which can be integrated to this genomic data to help understand the causes of frailty related diseases. DNA methylation data are generated on 850,000 methylation sites in 1,479 participants. The CLSA has also initiated a subcohort to collect longitudinal data from magnetic resonance imaging of the brain and microbiome of the gut in 6,000 participants. This data resource will facilitate the research on complex relationship between human genomic variants and a wide spectrum of environmental, lifestyle, and medical factors. The comprehensive pharmacogenomic and inflammation markers among other disease-associated variants may be of particular interest since DNA methylation and proteomic data are being generated. The CLSA overall is an ongoing perspective study. Follow-up data will continue to be collected from participants in the present genomic subcohort.

Collaboration

The genomic data from the CLSA Comprehensive cohort are accessible via the CLSA Data Access process (https://www.clsa-elcv.ca/data-access). The list of phenotypic variables can be browsed via the CLSA Data Preview Portal (https://datapreview.clsa-elcv.ca/). To be informed of the potential overlapping research topics, prospective data users are encouraged to consult the approved project summaries catalogued on the CLSA website (http://www.clsa-elcv.ca/researchers/approved-project-summaries). Given that this genomic data resource is released in 2018, we calculated the proportion of data requests including genomic data since 2018. At the time of writing, 17% of approved projects requested genetic data for their studies. The directly genotyped data are provided in binary PLINK format. It is recommended to use PLINK to manipulate these files (https://www.cog-genomics.org/plink/1.9/). The imputed genotyped data are provided in binary BGEN version 1.2 format using 8-bit encoding. It is recommended to use *qctool* version 2 or *bgenix* to manipulate this data type. The HLA imputation file is a plain text file containing information pertaining to the imputation of classical human leukocyte antigen alleles from SNP genotypes.

All studies using CLSA genetic data resource are required to give full acknowledgement to CLSA in their publications following instructions in *Publication and Promotion Policy for CLSA Data Users* on https:///www.clsa-elcv.ca.

Ethics statement

Ethics approval was provided by McMaster University Research Ethics Board. Study numbers: 10-423 2010-2336 11.003 C2010-80 2009-18 H10-02143 H2010:330 M16-10-023 2010s0527.

Funding

This research was made possible using the data collected by the Canadian Longitudinal Study on Aging (CLSA). Funding for the Canadian Longitudinal Study on Aging (CLSA) is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under

grant reference: LSA 94473 and the Canada Foundation for Innovation, as well as the following provinces (no award/grant number), Newfoundland, Nova Scotia, Quebec, Ontario, Manitoba, Alberta, and British Columbia. The CLSA is led by Drs. Parminder Raina, Christina Wolfson and Susan Kirkland. The work was also supported by Genome Canada Technology Platform #12505 and CFI#33408.

Author contributions

V.F. and R.L. conducted data analyses and drafted the manuscript, C.D-Z. and A.B. generated data, C.B., D.R., C.W., G.L., G.P., A.D.P., L.E.G., C.V., M.L., S.K., P.R., J.B.R., and J.R developed the concept and report design. All authors revised the manuscript critically for important intellectual content and approved the final version to be published.

Competing interests: None declared.

Data availability statement

Data are available from the Canadian Longitudinal Study on Aging (www.clsa-elcv.ca) for researchers who meet the criteria for access to de-identified CLSA data.

Reference:

1. Vineis P, Marinelli D, Autrup H, et al. Current smoking, occupation, N-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. *Cancer Epidemiol Biomarkers Prev* 2001;10(12):1249-52.
2. Wu C, Kraft P, Zhai K, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nat Genet* 2012;44(10):1090-7. doi: 10.1038/ng.2411 [published Online First: 20120909]
3. Singh PP, Demmitt BA, Nath RD, et al. The Genetics of Aging: A Vertebrate Perspective. *Cell* 2019;177(1):200-20. doi: 10.1016/j.cell.2019.02.038 [published Online First: 2019/03/23]
4. Melzer D, Pilling LC, Ferrucci L. The genetics of human ageing. *Nat Rev Genet* 2020;21(2):88-101. doi: 10.1038/s41576-019-0183-6 [published Online First: 2019/11/07]
5. Rask-Andersen M, Karlsson T, Ek WE, et al. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genet* 2017;13(9):e1006977. doi: 10.1371/journal.pgen.1006977 [published Online First: 20170905]
6. Raina P, Wolfson C, Kirkland S, et al. Cohort Profile: The Canadian Longitudinal Study on Aging (CLSA). *Int J Epidemiol* 2019;48(6):1752-53j. doi: 10.1093/ije/dyz173 [published Online First: 2019/10/22]
7. Affymetrix. UKB WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory. 2017. http://biobank.ndph.ox.ac.uk/showcase/docs/affy_data_generation2017.pdf.
8. Raina PS, Wolfson C, Kirkland SA, et al. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009;28(3):221-9. doi: 10.1017/S0714980809990055 [published Online First: 2009/10/29]
9. UK Biobank Axiom Array | UK Biobank [Available from: http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ accessed 10. Apr. 2018.
10. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26(22):2867-73. doi: 10.1093/bioinformatics/btq559 [published Online First: 2010/10/12]
11. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7. doi: 10.1186/s13742-015-0047-8 [published Online First: 2015/02/28]
12. Chang SPaC. PLINK 1.9 [Available from: https://www.cog-genomics.org/plink1.9 accessed 27. Apr 2018.
13. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526(7571):68-74. doi: 10.1038/nature15393

14. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005;76(5):887-93. doi: 10.1086/429864 [published Online First: 2005/03/25]
15. Galinsky KJ, Bhatia G, Loh PR, et al. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 2016;98(3):456-72. doi: 10.1016/j.ajhg.2015.12.022 [published Online First: 2016/03/01]
16. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7(10):781-91. doi: 10.1038/nrg1916 [published Online First: 2006/09/20]
17. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38(8):904-9. doi: 10.1038/ng1847 [published Online First: 2006/07/25]
18. Dilthey A, Leslie S, Moutsianas L, et al. Multi-population classical HLA type imputation. *PLoS Comput Biol* 2013;9(2):e1002877. doi: 10.1371/journal.pcbi.1002877 [published Online First: 2013/03/06]
19. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* 2015;31(5):782-4. doi: 10.1093/bioinformatics/btu704 [published Online First: 2014/10/24]
20. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;590(7845):290-99. doi: 10.1038/s41586-021-03205-y [published Online First: 2021/02/12]
21. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-87. doi: 10.1038/ng.3656 [published Online First: 2016/08/30]
22. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48(11):1443-48. doi: 10.1038/ng.3679 [published Online First: 2016/10/28]
23. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-83. doi: 10.1038/ng.3643 [published Online First: 2016/08/23]
24. Blindness GBD, Vision Impairment C, Vision Loss Expert Group of the Global Burden of Disease S. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* 2021;9(2):e144-e60. doi: 10.1016/S2214-109X(20)30489-7 [published Online First: 20201201]
25. Han X, Steven K, Qassim A, et al. Automated AI labeling of optic nerve head enables insights into cross-ancestry glaucoma risk and genetic discovery in >280,000 images from UKB and CLSA. *Am J Hum Genet* 2021;108(7):1204-16. doi: 10.1016/j.ajhg.2021.05.005 [published Online First: 20210601]
26. Forgetta V, Keller-Baruch J, Forest M, et al. Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study. *PLoS Med* 2020;17(7):e1003152. doi: 10.1371/journal.pmed.1003152 [published Online First: 2020/07/03]

27. Menniti G, Paquet C, Han HY, et al. Multiscale Risk Factors of Cardiovascular Disease: CLSA Analysis of Genetic and Psychosocial Factors. *Front Cardiovasc Med* 2021;8:599671. doi: 10.3389/fcvm.2021.599671 [published Online First: 2021/04/03]
28. Campos AI, Ingold N, Huang Y, et al. Genome-wide analyses in 1,987,836 participants identify 39 genetic loci associated with sleep apnoea. *medRxiv* 2020:2020.09.29.20199893. doi: 10.1101/2020.09.29.20199893
29. UK Biobank Axiom Array 2017 [Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fbrochures%2Fuk_axiom_biobank_genotyping_arrays_datasheet.pdf.
30. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562(7726):203-09. doi: 10.1038/s41586-018-0579-z [published Online First: 20181010]

Table1: Count of CLSA genotyped participants by self-reported sex and sex chromosome composition

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

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

Inferred Relationship	Count
Monozygotic twin	1
Full sibling	357
Parent/offspring	176
2nd degree	315
3rd degree	1066
Unrelated	123294

Table 3: Count of CLSA genotyped participants per self-reported ancestry and k-means cluster

Self-reported ancestry[a]	k-means cluster					
	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

[a]The details of grouping self-reported cultural and racial category into fewer groups are in Supplementary Table S2

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

The SNPs are considered as passed if they have nonsignificant p-value (Fisher's $p > 3.5 \times 10^{-10}$) below the multiple testing corrected threshold for the respective test on discordant genotype frequency between batch, departure from HWE, discordance between the positive control replicates and on discordant genotype frequency between male and female.

Figure 2: Determining the CLSA European ancestry subset.

(A) Top 4 principal components from all 1000 Genomes populations labelled and coloured. Population code refers to https://www.internationalgenome.org/category/population/. (B) Top 4 principal components from CLSA color coded and labelled by cluster number.

Figure 3: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X and Y are shown for sex chromosome abnormalities.

(A) In 46,XY, the BAF is either 0 or 1 and the expected log2 Ratio is less than 0 on chromosome X. However, in the pseudoautosomal region (PAR) and the chrY11.2/chrXq21.3 homology block, there are heterozygous calls in male shown as BAF of 0.5. The red line shows the lowess curve for log2 Ratio. The BAF is either 0 or 1 and the expected log2 Ratio is 0 on chromosome Y. (B) In 46,XX, the BAF is either 0 (AA), ½ (AB) or 1 (BB) and the expected Log2 Ratio is 0 on chromosome X as in a normal diploid cell. The BAF is between 0 and 1, and Log2 Ratio is less than 0 on chromosome Y. (C) For Klinefelter syndrome (47,XXY), log2 ratio is around 0 on chromosome X which indicates ploidy as 2N. Compared to 46,XY, there is relatively lower peaks of log2 ratio at PAR and chrX21.3/chrY11.2 homology block region. And BAF of heterozygous calls at PAR and chrX21.3/chrY11.2 homology block region shifted from 0.5 to intermediate values. They both indicated an extra copy of chromosome X. Chromosome Y intensity profile showed clear male pattern. (D) For Turner syndrome (45,X), on chromosome X, log2 ratio is below 0 and there is no BAF bands of 0.5, which indicates one copy loss. Chromosome Y intensity profile showed clear female pattern. (E) For 45,X/46,XX mosaicism, on chromosome X, there is a relatively smaller decrease of log2 ratio compared to 1 copy of chromosome X as in male. The BAF of heterozygous calls on chromosome X is split to intermediate values. They both indicate that the sample is mosaic for deletion of chromosome X. Chromosome Y intensity profile showed clear female pattern. (F) For 45,X/46,XY mosaicism, the log2 ratio less than 0 and no BAF 0.5 band on chromosome X indicates one copy. The log2 ratio shifts to below 0 and BAF values between 0 and 1 on chromosome Y indicates chromosome loss. However, the intermediate BAF values close to 0 or 1 at PAR and chrX21.3/chrY11.2 homology block region indicates the loss of chromosome Y is existed in a larger proportion of cells.

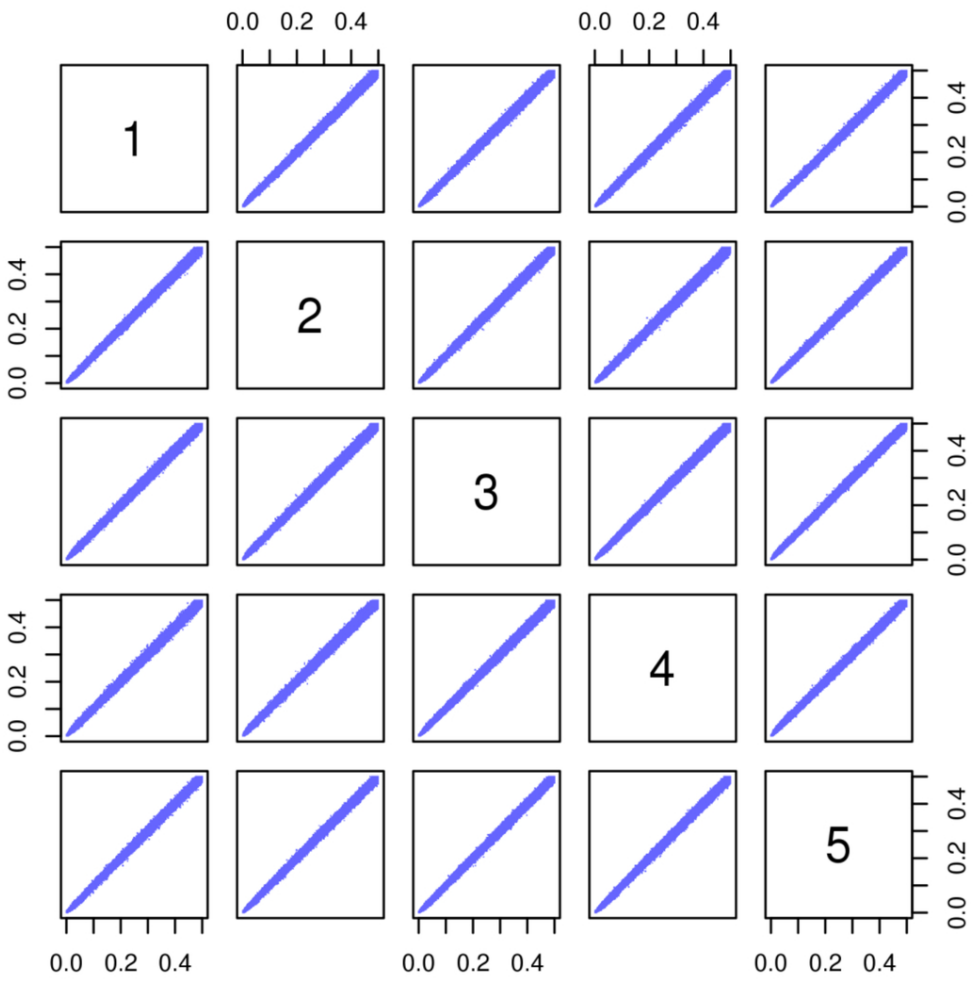


Figure 1: Pairwise plot of allele frequency of SNPs that pass all 4 tests from genotype batch 1 to 5. The SNPs are considered as passed if they have nonsignificant p-value (Fisher's $p > 3.5 \times 10^{-10}$) below the multiple testing corrected threshold for the respective test on discordant genotype frequency between batch, departure from HWE, discordance between the positive control replicates and on discordant genotype frequency between male and female.

89x88mm (300 x 300 DPI)

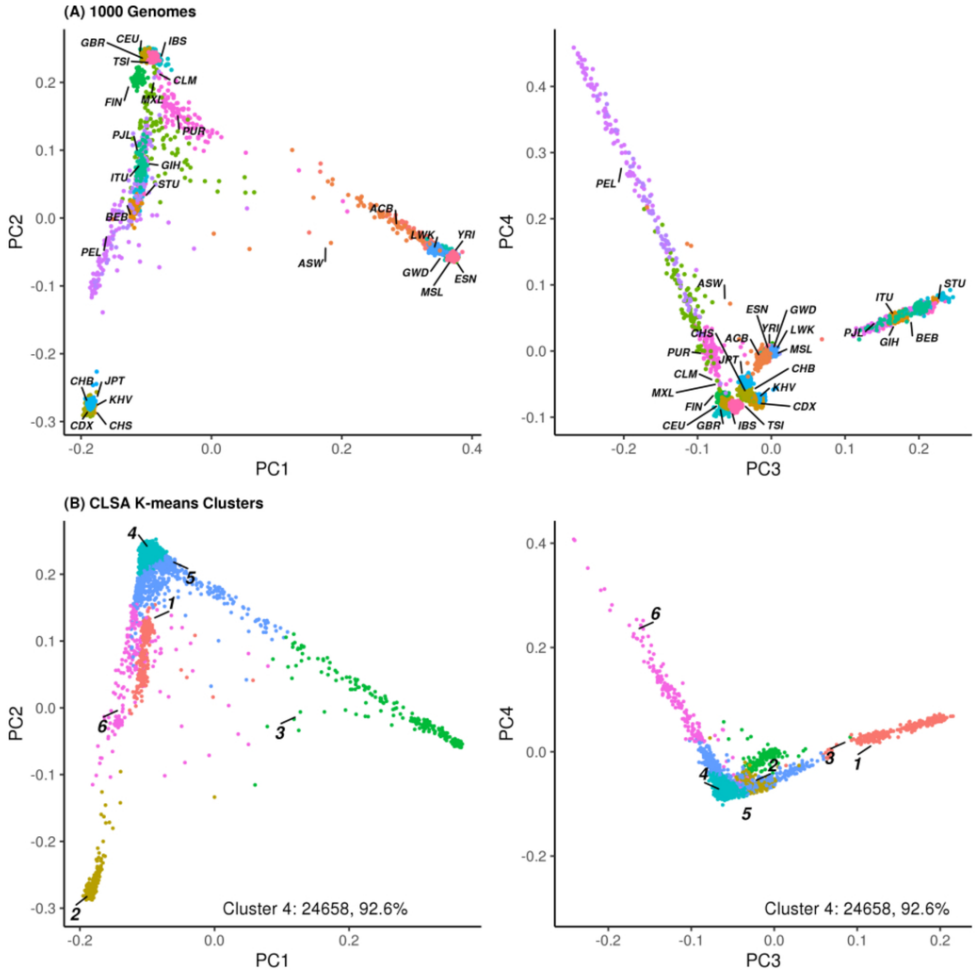


Figure 2: Determining the CLSA European ancestry subset.
(A) Top 4 principal components from all 1000 Genomes populations labelled and coloured. Population code refers to https://www.internationalgenome.org/category/population/. (B) Top 4 principal components from CLSA color coded and labelled by cluster number.

89x89mm (300 x 300 DPI)

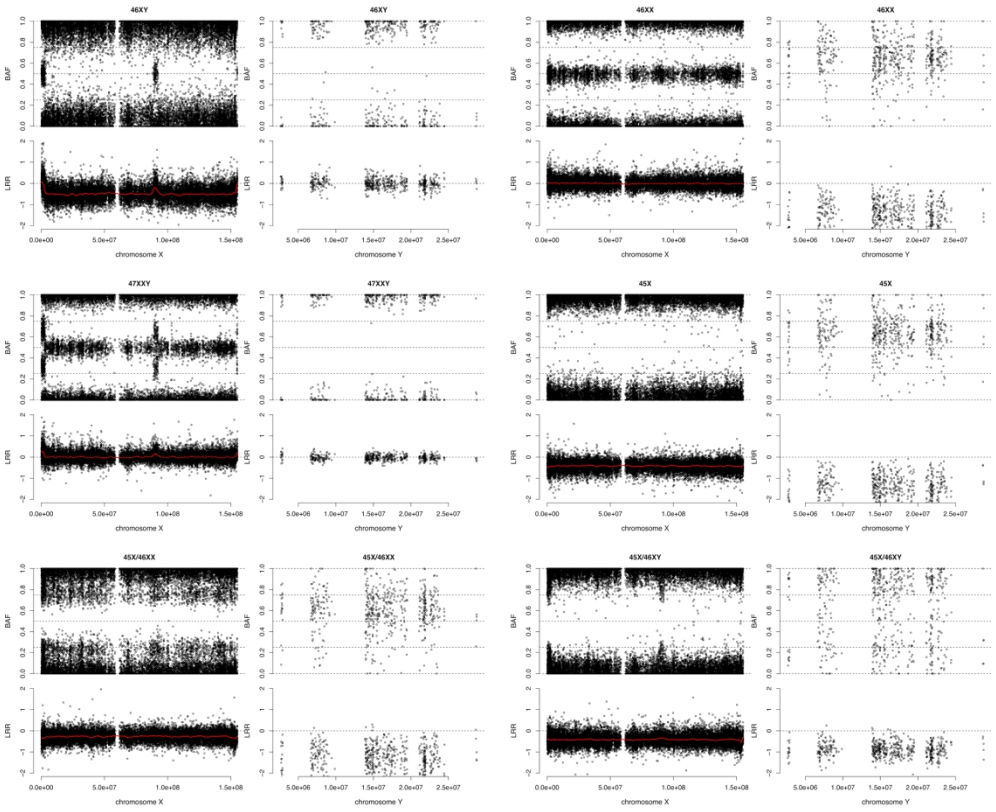


Figure 3: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X and Y are shown for sex chromosome abnormalities.

(A) In 46,XY, the BAF is either 0 or 1 and the expected log2 Ratio is less than 0 on chromosome X. However, in the pseudoautosomal region (PAR) and the chrY11.2/chrXq21.3 homology block, there are heterozygous calls in male shown as BAF of 0.5. The red line shows the lowess curve for log2 Ratio. The BAF is either 0 or 1 and the expected log2 Ratio is 0 on chromosome Y. (B) In 46,XX, the BAF is either 0 (AA), ½ (AB) or 1 (BB) and the expected Log2 Ratio is 0 on chromosome X as in a normal diploid cell. The BAF is between 0 and 1, and Log2 Ratio is less than 0 on chromosome Y. (C) For Klinefelter syndrome (47,XXY), log2 ratio is around 0 on chromosome X which indicates ploidy as 2N. Compared to 46,XY, there is relatively lower peaks of log2 ratio at PAR and chrX21.3/chrY11.2 homology block region. And BAF of heterozygous calls at PAR and chrX21.3/chrY11.2 homology block region shifted from 0.5 to intermediate values. They both indicated an extra copy of chromosome X. Chromosome Y intensity profile showed clear male pattern. (D) For Turner syndrome (45,X), on chromosome X, log2 ratio is below 0 and there is no BAF bands of 0.5, which indicates one copy loss. Chromosome Y intensity profile showed clear female pattern. (E) For 45,X/46,XX mosaicism, on chromosome X, there is a relatively smaller decrease of log2 ratio compared to 1 copy of chromosome X as in male. The BAF of heterozygous calls on chromosome X is split to intermediate values. They both indicate that the sample is mosaic for deletion of chromosome X. Chromosome Y intensity profile showed clear female pattern. (F) For 45,X/46,XY mosaicism, the log2 ratio less than 0 and no BAF 0.5 band on chromosome X indicates one copy. The log2 ratio shifts to below 0 and BAF values between 0 and 1 on chromosome Y indicates chromosome loss. However, the intermediate BAF values close to 0 or 1 at PAR and chrX21.3/chrY11.2 homology block region indicates the loss of chromosome Y is existed in a larger proportion of cells.

90x72mm (600 x 600 DPI)

Supplementary Table S1: Sex chromosome determination of miscalled genotyped CLSA participants

Self reported sex	Affymetrix sex corrected by SVM	PLINK sex (raw F estimate <0.3 female raw F estimate >0.8 male)	discordance in section Sex chromosome composition	PLINK sex (adjusted F estimate <0.4 female adjusted F estimate >0.8 male)	sex determined by combined Affymetrix/PLINK/CNV approach	CLSA self-reported phenotype	chromosomal abnormality from CNV profile	Raw F estimate	Adjusted F estimate
male	female	female	1	female	female		No abnormality	-0.00301	0.01745
male	male	female	1	female	male		Klinefelter Syndrome (47,XXY)	0.1048	0.1002
female	male	male	1	male	male		No abnormality	1	1
male	female	female	1	female	female		No abnormality	0.01533	0.02791
female	female	unknown		female	female		Low heterozygosity on chrX	0.3495	0.3644
female	female	unknown	1	unknown	female		45,X/46,XX mosaicism	0.5095	0.4773
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.4637	0.4485
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.4352	0.4746
male	female	female	1	female	female		No abnormality	0.008892	0.03223
female	male	male	1	male	male		No abnormality	1	1
female	male	male	1	male	male		No abnormality	1	1
male	female	female	1	female	female		No abnormality	-0.03065	-0.02794
female	female	male	1	male	female	Turner Syndrome	Turner Syndrome (45,X)	0.9507	0.9614
female	male	male	1	male	male		No abnormality	1	1
female	female	unknown		female	female		Low heterozygosity on chrX	0.3043	0.2802
male	female	female	1	female	female		No abnormality	-0.002689	-0.01603
female	male	male	1	male	male		No abnormality	1	1
male	male	unknown	1	female	male		Klinefelter Syndrome (47,XXY)	0.3124	0.3235
female	male	male	1	male	male		No abnormality	1	1
male	male	unknown	1	unknown	male		No abnormality	0.6854	0.6878
male	male	unknown	1	unknown	male		45,X/46,XY mosaicism	0.5798	0.5893
female	female	male	1	male	female		Turner Syndrome (45,X)	0.9792	0.9875
female	male	male	1	male	male		No abnormality	1	1
male	female	female	1	female	female		No abnormality	0.04474	0.03819
female	female	male	1	male	female		Turner Syndrome (45,X)	0.9439	0.9545
male	male	unknown		male	male		45,X/46,XY mosaicism	0.7965	0.8497
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.4524	0.4305
male	female	male		male	female		45,X/46,XY mosaicism	0.9748	0.986
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.5457	0.5504

male	female	male		male	male		45,X/46,XY mosaicism	0.9678	0.9797
female	female	male	1	unknown	female		Low heterozygosity on chrX	0.8076	0.7966
female	male	male	1	male	male		No abnormality	1	1
female	male	unknown	1	male	female	Turner Syndrome	45,X/46,XY mosaicism	0.7827	0.8404
male	male	female	1	female	male		Klinefelter Syndrome (47,XXY)	-0.0457	-0.04304
female	female	male	1	male	female		Turner Syndrome (45,X)	0.879	0.9028
female	female	unknown		female	female		Low heterozygosity on chrX	0.3035	0.3339
male	male	unknown	1	unknown	male		45,X/46,XY mosaicism	0.6732	0.7016
male	male	female	1	female	male	Klinefelter Syndrome	Klinefelter Syndrome (47,XXY)	-0.03886	-0.02651
female	male	male	1	male	male		No abnormality	1	1
male	male	unknown	1	female	male		Klinefelter Syndrome (47,XXY)	0.3093	0.2872
male	female	male		male	male		45,X/46,XY mosaicism	1	0.9603
female	female	male	1	male	female	Turner Syndrome	Turner Syndrome (45,X)	0.9273	0.9527
female	female	unknown		female	female		Low heterozygosity on chrX	0.4004	0.3584
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.3838	0.4362
male	male	unknown		male	male		45,X/46,XY mosaicism	0.7707	0.8358
female	female	unknown		female	female		Low heterozygosity on chrX	0.3978	0.3886
female	female	unknown	1	unknown	female		Low heterozygosity on chrX	0.7748	0.7842
female	female	unknown		female	female		No abnormality	0.3298	0.3527
male	male	unknown	1	unknown	male		45,X/46,XY mosaicism	0.6658	0.7601
male	female	male		male	male		45,X/46,XY mosaicism	1	0.9827
female	female	unknown	1	unknown	female		45,X/46,XX mosaicism	0.4148	0.4228
male	female	female	1	female	female		No abnormality	-0.003668	-0.01753
female	female	unknown		female	female		Low heterozygosity on chrX	0.3489	0.2975
female	male	male	1	male	male		No abnormality	1	1
male	female	female	1	female	female		No abnormality	-0.02345	-0.01932
male	male	female	1	female	male	Klinefelter Syndrome	Klinefelter Syndrome (47,XXY)	-0.03581	-0.02924
female	female	unknown	1	unknown	female		45,X/46,XX mosaicism	0.5139	0.5022
female	female	unknown		female	female		Low heterozygosity on chrX	0.3461	0.3336
male	female	female	1	female	female		No abnormality	0.01936	-0.006017
male	female	female	1	female	female		No abnormality	-0.04206	-0.0411
female	male	male	1	male	male		No abnormality	1	1
male	female	male		male	male		45,X/46,XY mosaicism	0.9756	0.9832

male	male	unknown	1	unknown	male	45,X/46,XY mosaicism	0.72	0.7971	

Supplementary Table S2: Self-reported ancestry and derived category from cultural and racial background

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
Mixed	Mixed
Other	Other
Refused	Other
South Asian	South Asian
Filipino	Southeast Asian
Southeast Asian	Southeast Asian
White	White

Supplementary Table S3: Comparison of HLA types in positive controls with known types in literature

HLA locus	NA10859 (1347-02)[a] Reference Genotype	NA10859 (1347-02) No of replicates	NA10859 (1347-02) Accuracy(%)/ Call rate(%)[d]	NA12878 (1463-02) Reference Genotype	NA12878 (1463-02) No of replicates	NA12878 (1463-02) Accuracy(%)/ Call rate(%)	NA24385 (HG002 (NIST RM 8391)) Reference Genotype	NA24385 (HG002 (NIST RM 8391)) No of replicates	NA24385 (HG002 (NIST RM 8391)) Accuracy(%)/ Call rate(%)
A	03:01/01:01[b]	4	100/100	01:01/11:01[e]	587	100/100	26:01/01:01[f]	75	100/100
B	07:02/15:01[b]	4	100/100	08:01/56:01[e]	587	100/100	38:01/35:08[f]	75	-
C	06:02/07:02[b]	4	100/100	01:02/07:01[e]	587	100/100	12:03/04:01[f]	75	100/100
DPA1	01:03/01:03[b]	4	100/100	01:03/02:01[e]	587	100/100	01:03/01:03[f]	75	100/100
DPB1	04:02/04:02[b]	4	100/100	04:01/14:01[e]	587	100/100	04:01/04:01[g]	75	100/100
DQA1	01:02/03:01[b]	4	100/100	01:01/05:01[e]	587	100/100	03:01/01:01[h]	75	100/100
DQB1	03:02/06:02[b]	4	100/100	02:01/05:01[e]	587	100/100	05:01/03:02[f]	75	100/100
DRB1	04:01/15:01[b]	4	100/100	01:01/03:01[e]	587	100/100	04:02/10:01[f]	75	100/100
DRB3	NA[b,c]	4	-	01:01/01:01;01:01/02:02[e]	587	100/100	NA[f]	75	-
DRB4	01[b]	4	100/100	01:01/01:01;01:03/01:03;01:06/01:06[e]	587	-	01:03[f]	75	100/100
DRB5	Na[b,c]	4	-	NA[e]	587	-	NA[f]	75	-

Note: a-Coriell ID (CEPH Family ID or NIST ID/RM Number for Personal Genome Project sample)

b: reference genotype source-IPD-IMGT/HLA Database

c: reference genotype data is not available

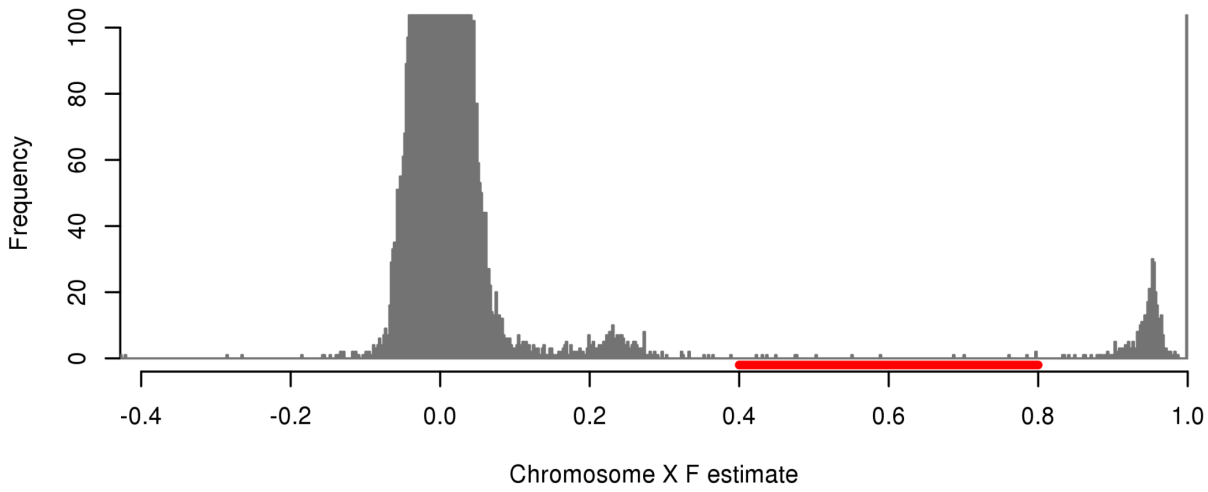
d: call rate is based on a posterior probability call threshold of 0.7

e: reference genotype source-PLoS Comput Biol. 2016 Oct; 12(10): e1005151. PMID: 27792722. A set of possible alleles are reported in the reference. The HLA types we validated are shown in the table.

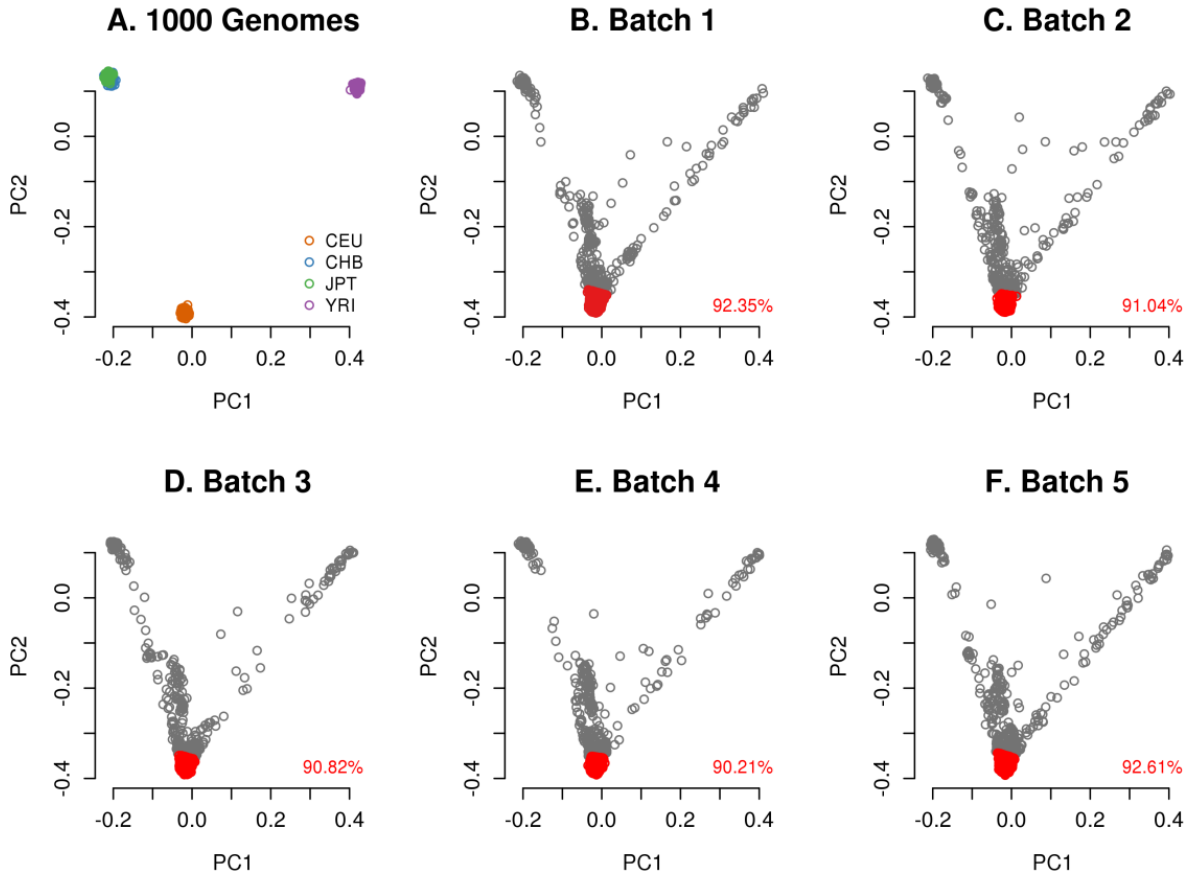
f: reference genotype source-DOI: 10.12688/f1000research.19630.1

g: reference genotype source-DOI: 10.12688/f1000research.19630.1 and https://www.pacb.com/wp-content/uploads/Rowell-CSHLBioData-2018-Comprehensive-Variant-Detection-in-a-Human-Genome-with-PacBio-High-Fidelity-Reads.pdf. The HLA types we validated are shown in the table.

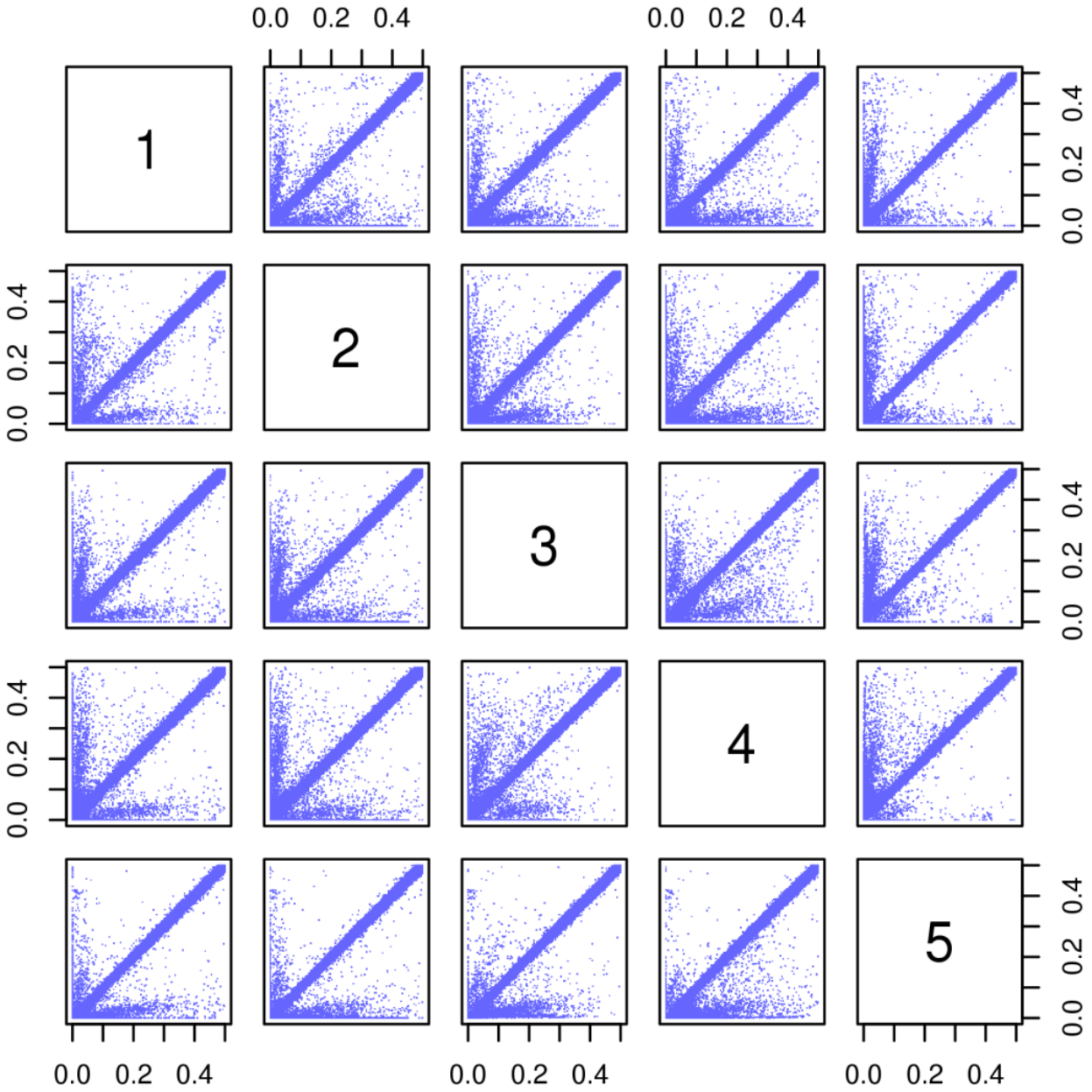
h: reference genotype source-DOI: 10.12688/f1000research.19630.1 and Nature Communications doi: 10.1038/s41467-020-18564-9. The HLA types we validated are shown in the table.



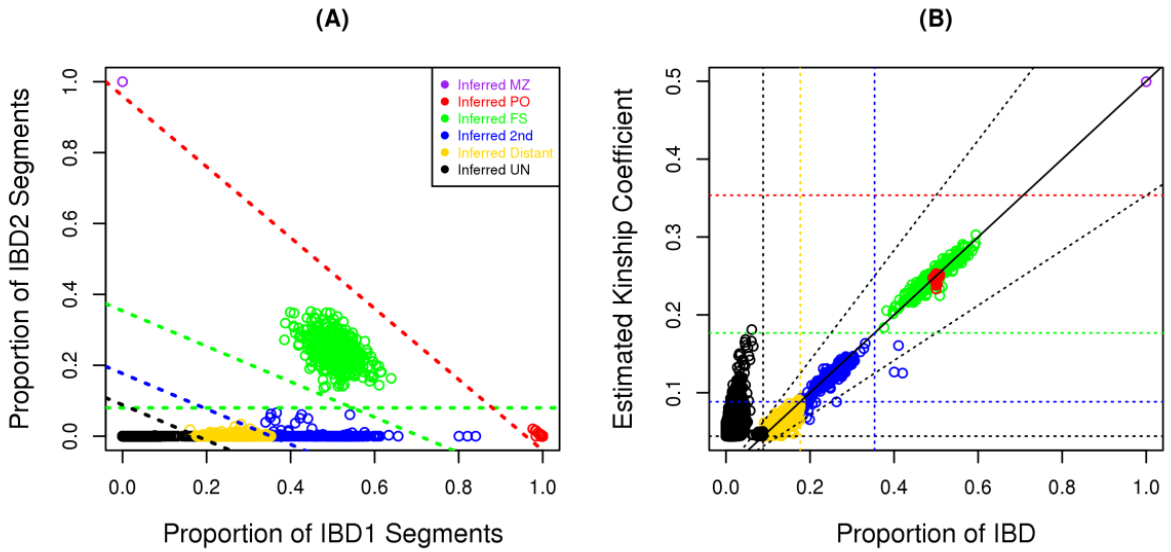
Supplementary Figure S1: Distribution of chromosome X F estimates for CLSA genotyped participants (y-axis truncated). Individuals with chromosome X F estimates within the range of 0.4 to 0.8 (red) are considered to have undefined chromosomal sex.



Supplementary Figure S2: 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).

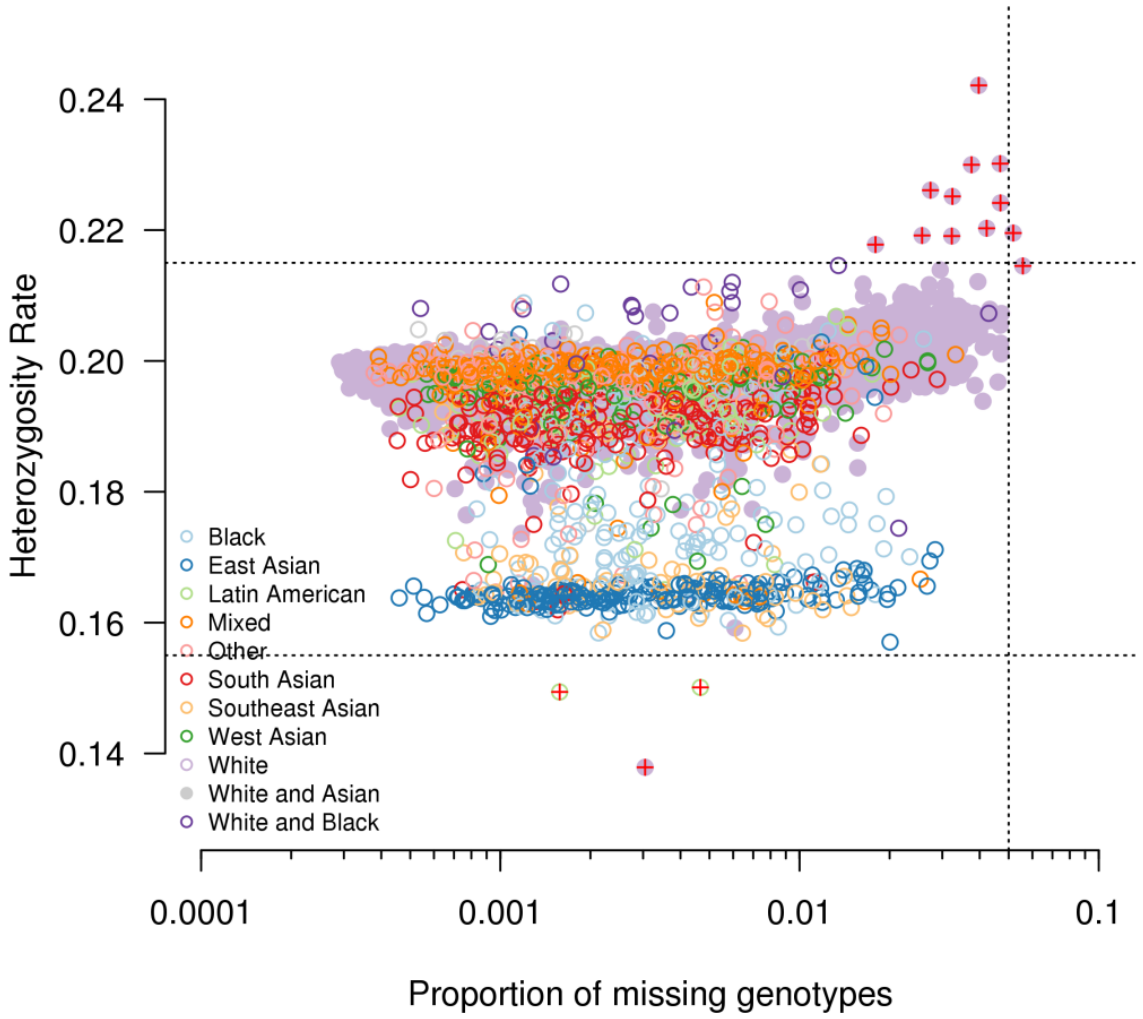


Supplementary Figure S3: Pairwise plot of allele frequency of SNPs from genotype batch 1 to 5.

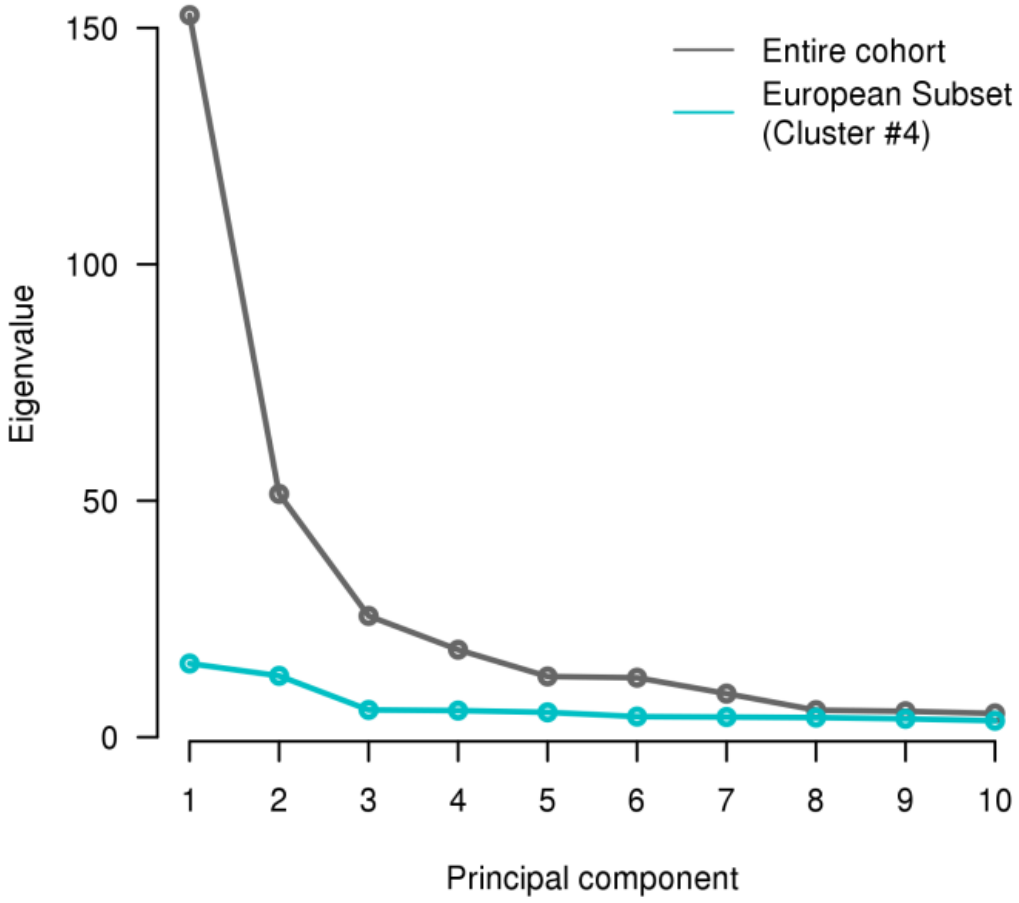


Supplementary Figure S4: Inference of familial relatedness using KING.

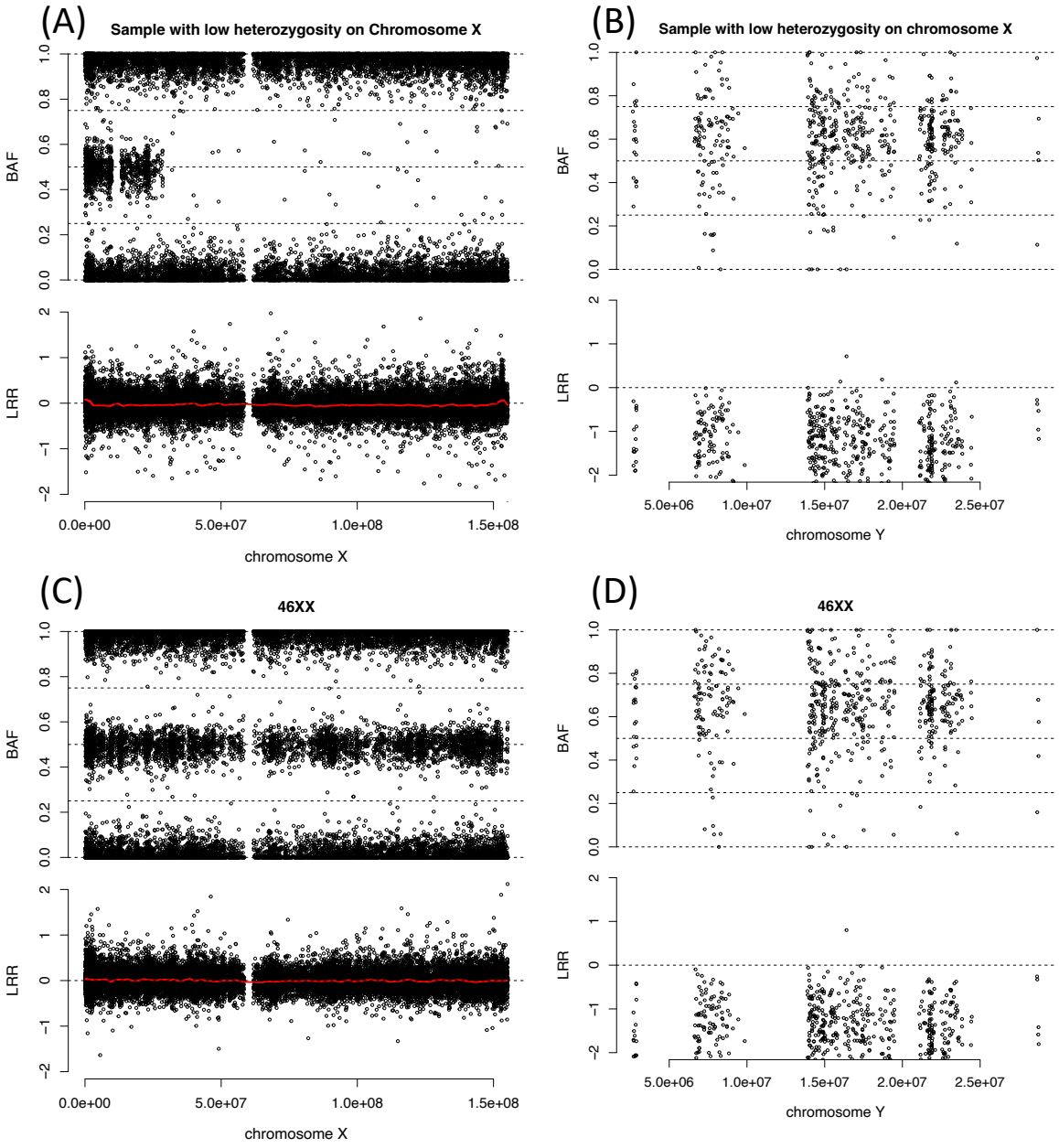
(A) Inference using IBD segments. (B) Inference using proportion IBD and kinship coefficient. Relationships in legend are abbreviated as: MZ=Monozygotic twin, PO=Parent/offspring, FS=Full sibling, 2nd=Second-degree relative, 3rd=Third-degree relative, Distant=Greater than 3rd degree relative, UN=Unrelated.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



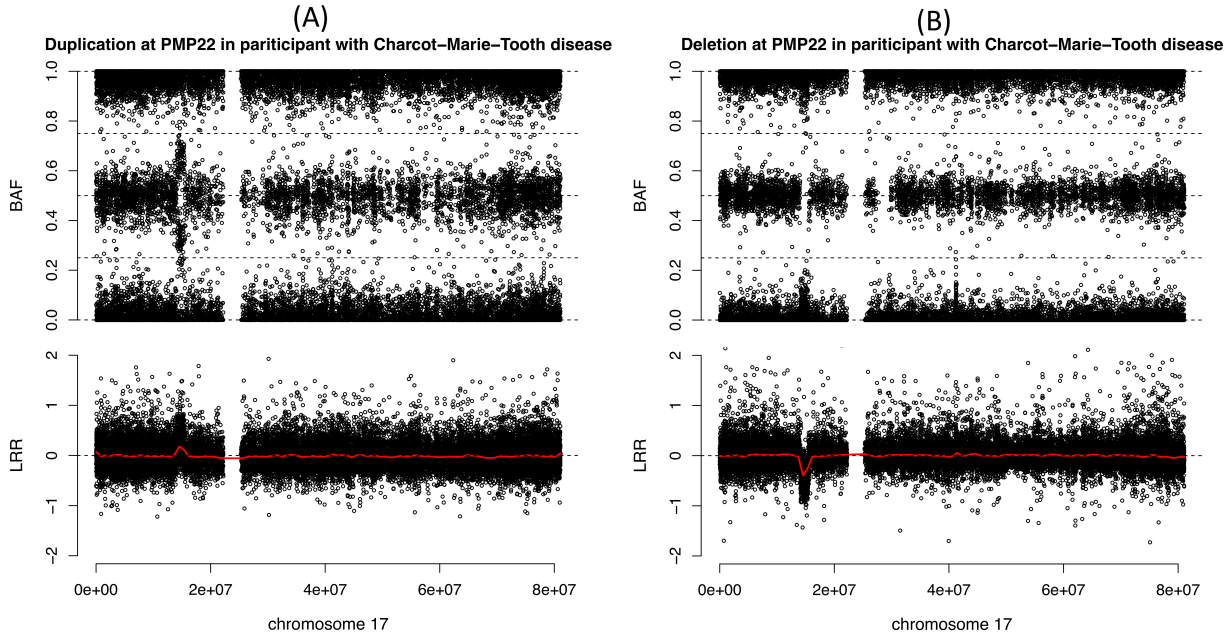
Supplementary Figure S5: Sample-wise heterozygosity versus genotype missingness. Points are color coded according to self-reported ancestry category. Outliers are marked with a red plus sign.



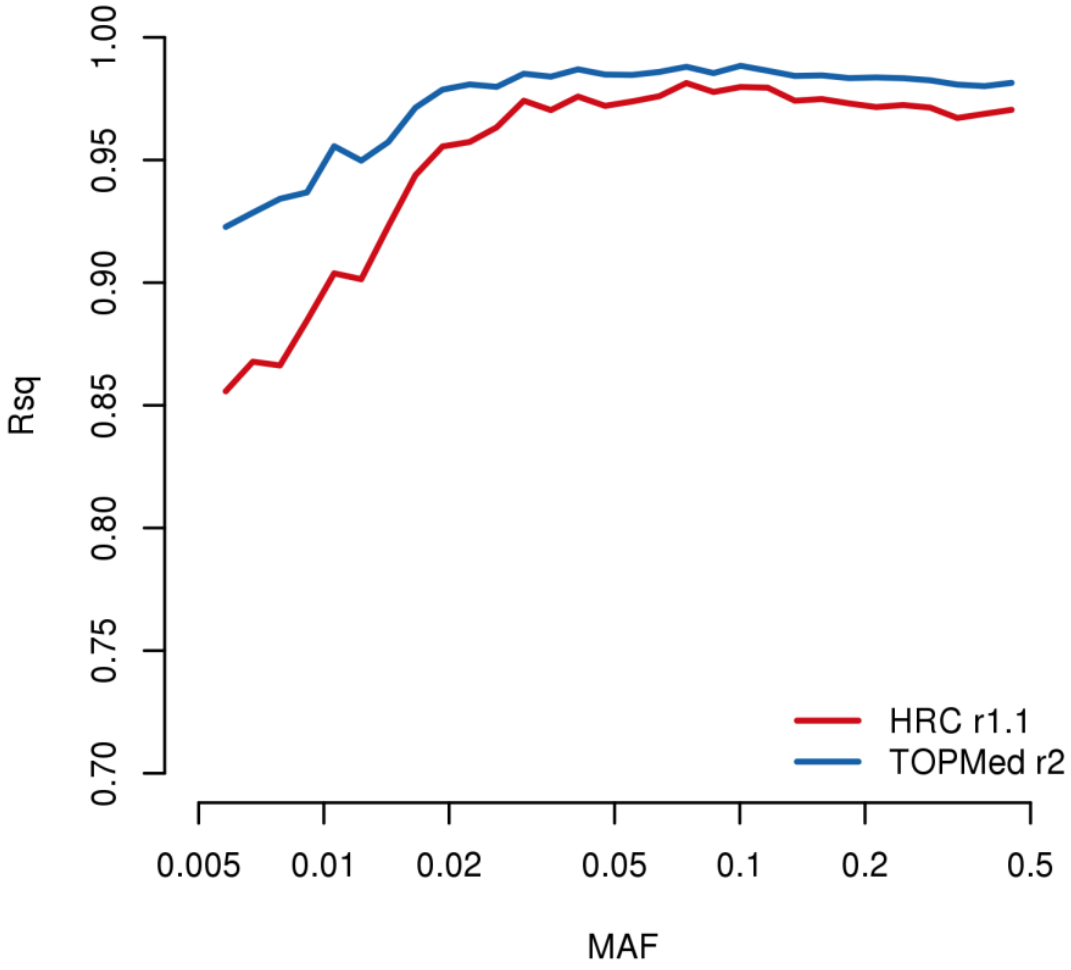
Supplementary Figure S6: Eigenvalues for PCA analysis of the entire cohort (grey) and the European ancestry subset (cluster 4, Robin egg blue), demonstrating a reduction in genetic variance within the European ancestry subset.



Supplementary Figure S7: BAF (TOP) and log2 ratio (BOTTOM) of chromosomes X (A) and Y (B) are shown for sample with low heterozygosity on chromosome X compared to sample with 46,XX (C-D).



Supplementary Figure S8: BAF (TOP) and log2 ratio (BOTTOM) of chromosome 17 are shown for sample with duplication (A) or deletion (B) at *PMP22* locus.



Supplementary Figure S9: Imputation quality of the CLSA cohort using the TOPMed versus Haplotype Reference Consortium (HRC) reference panel stratified by minor allele frequency (MAF) bins (data shown is from chromosome 22).