BMJ Open

Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA)

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

To cite: Forgetta V, Li R, Darmond-Zwaig C, *et al*. Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). *BMJ Open* 2022;**12**:e059021. doi:10.1136/bmjopen-2021-059021

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-059021).

VF and RL are joint first authors. PR, JBR and JR are joint senior authors.

Received 05 November 2021
Accepted 08 February 2022

For numbered affiliations see end of article.

Correspondence to
Dr Jiannis Ragoussis;
ioannis.ragoussis@mcgill.ca

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 ageing process. The aim of this report was to describe the genomic data included in CLSA.

Participants A total of 26 622 individuals from the CLSA Comprehensive cohort of men and women aged 45–85 recruited between 2010 and 2015 underwent 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 human leukocyte antigen gene 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. These genomic data were 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 on 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 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 an 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 200 years. In such times, the make-up of Canadian population has changed unprecedentedly. From 1977 to 2017, the senior population, that is, people aged 65 years and older, grew from 2.0 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 four 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 healthcare systems inadvertently. Age-related diseases have a highly complex nature. Both genetic and environmental factors play an important role as well as the interaction between them.[1 2] Therefore, understanding of the underlying mechanisms of ageing 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 ageing 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. Genome-wide association studies (GWASs) have uncovered tens of genes and genetic variations that play a role in the variability of ageing 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 were 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 were 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

There was no patient and public involvement in this cohort profile.

DATA COLLECTED

Sample storage and DNA extraction

The biological samples were collected at the data collection sites and were deidentified.[8] 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 1 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 2 hours for all participants. Genomic DNA was extracted from blood samples using the purification protocol 'Chemagic DNA Buffy Coat Kit Special 200 µL Prefilling VD151007' on the Chemagic MSM I instrument (article no. CMG-533; PerkinElmer, Baesweiler, Germany). All extracted samples were quantified using PicoGreen Reagent Kit (catalogue # P7589, Life Technologies). A minimum DNA concentration for passing of samples was set at 10 ng/µL. Samples were subsequently normalised 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 four controls, one male control as the Affymetrix Reference Genomic DNA 103 (catalogue # 900421) or Personal Genome Project sample huAA53E0 (Coriell Cell Repositories, catalogue # NA24385), two female controls as the Centre d'Etudes du Polymorphisme Humain (CEPH) control 1463–02 (Coriell Cell Repositories, catalogue # NA12878) or the CEPH control 1347–2 (Coriell Cell Repositories, catalogue # NA10859), and a deionised water negative control. The Affymetrix protocol (Axiom 2.0 Assay Automated Workflow on Affymetrix NIMBUS) was followed. Samples were hybridised to UK Biobank arrays (ThermoFisher, catalogue #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 eight plates were analysed using the sample QC workflow of the Axiom Analysis Suite V.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 QC and removal of duplicate genotyped participants

Genotyping was undertaken in separate batches of approximately 5000 samples each using Axiom Analysis Suite V.2.0, similar to UK Biobank genotyping QC

Table 1 Count of Canadian Longitudinal Study on Aging genotyped participants by self-reported sex and sex chromosome composition

Self-reported sex	Sex chromosome composition	Count
Male	Male	13 324
Female	Female	13 250
Female	Male	17
Male	Female	16
Female	Undefined	10
Male	Undefined	5

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 of ≥0.82 on sample level, average QC call rate of passing samples on a plate (plate QC call rate) of ≥95%, percentage of passing samples of ≥70% and average call rate for passing samples of ≥95% on plate level. Duplicate genotyped participants were detected by KING V.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 (online supplemental figure S1). Using this threshold, we obtained X chromosome number using PLINK V.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 online supplemental table S1. All subsequent analyses in this paper will use X chromosome number and number of non-missing Y chromosome genotypes to define sex.

Genetic marker-based QC

This consisted of four 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 aforementioned 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 four 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 III.[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) (online supplemental figure S2).

We then set a multiple-testing corrected p value threshold for QC tests as 3.15×10^{-10}. For the 794 409 markers and five 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 QC metrics that satisfy their research requirements.

Discordant genotype frequency between batches

To detect deviation in genotype frequency of markers between batches, we used Fisher's exact test on the 2×3 table of genotype counts (or 2×2 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 HWE

We conducted the test for departure from HWE using the exact test.[14] There were 7790 markers with an HWE p value of $<3.15\times10^{-10}$.

Discordance across control replicates

There were three 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 follows:

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

where n_{aa}, n_{ab} and n_{bb} are 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 (ie, 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 2×3 table of genotype counts for autosomal SNPs (or 2×2 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 of $<3.15\times10^{-10}$, 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 four tests. They are labelled in the marker QC file accompanying this data release. The effect of this quality analysis is depicted by comparing online supplemental figure S3 with figure 1 where there is clear improvement in the concordance in minor allele frequency

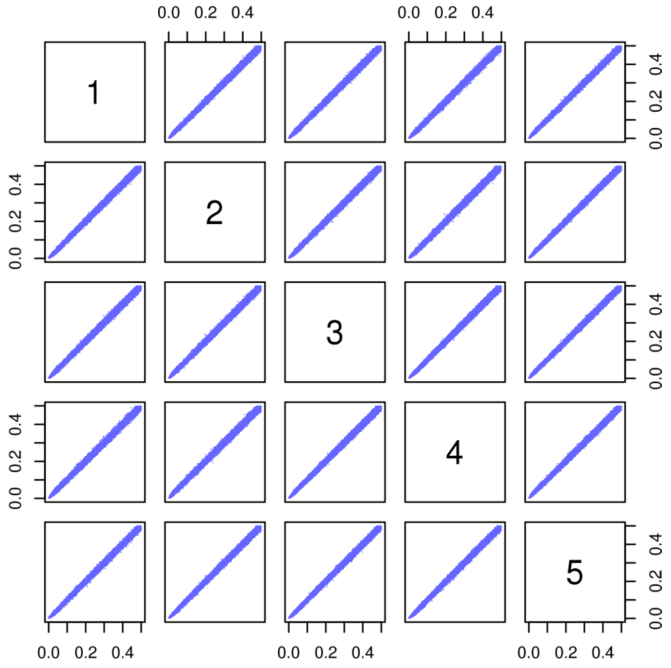

Figure 1 Pairwise plot of allele frequency of SNPs that pass all four tests from genotype batches 1–5. The SNPs are considered as passed if they have non-significant p value (Fisher's $p>3.5\times10^{-10}$) below the multiple testing corrected threshold for the respective test on discordant genotype frequency between batch, departure from Hardy-Weinberg equilibrium, discordance between the positive control replicates and on discordant genotype frequency between male and female. SNP, single-nucleotide polymorphism

(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 of <0.005 were flagged as they may bias subsequent sample-based QC.

Sample-based QC

This sample-based QC was intended to identify samples of low-quality, related individuals and to 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 QC metrics that satisfy their research requirements.

We selected the SNPs that passed all four tests from marker-based QC with MAF of >0.01 and marker-wise missingness of <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 a 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 KING Software,[10] we computed all pairwise kinship coefficients and noted all pairs with inferred relatedness of third degree or closer using autosomal SNPs (table 2 and online supplemental figure S4). Individuals with an inferred relationship of third degree or closer are labelled in the database.

Table 2 Count of kinship pairs per type of inferred relationship

Inferred relationship	Count
Monozygotic twin	1
Full sibling	357
Parent/offspring	176
Second degree	315
Third degree	1066
Unrelated	123 294

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 (online supplemental 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 third 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 four principal components from the analysis of population structure in the previous section into six clusters. Visualisation 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

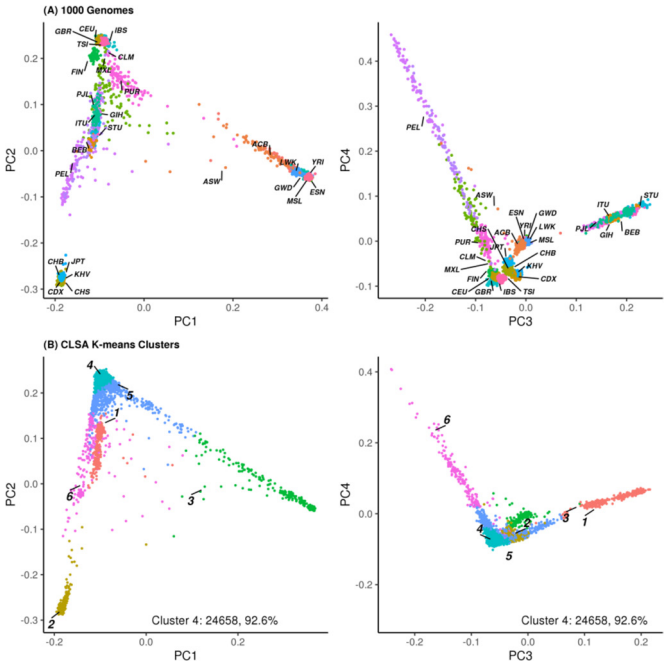

Figure 2 Determining the CLSA European ancestry subset. (A) Top four principal components from all 1000 Genomes populations labelled and coloured (population code refers to https://wwwinternationalgenomeorg/category/population/). (B) Top four principal components from CLSA colour coded and labelled by cluster number. CLSA, Canadian Longitudinal Study on Aging.

individuals in CLSA that self-report European ancestry (table 3 and online supplemental table S2). The European ancestry subset has markedly reduced variance in the top principal components as compared with the entire CLSA cohort (online supplemental 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.

Table 3 Count of Canadian Longitudinal Study on Aging genotyped participants per self-reported ancestry and k-means cluster

Self-reported ancestry*	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	24 380	742	41
White and Asian	3	3	0	5	19	11
White and black	2	0	11	3	17	0

*The details of grouping self-reported cultural and racial category into fewer groups are in online supplemental table S2.

Detection of copy number abnormalities associated with disease

Sex chromosome abnormalities

The sex chromosome composition was called by both Affymetrix Axiom Analysis Suite V.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 men by stringent Affymetrix default threshold, the intensity data of chromosome X and Y markers from all UK Biobank samples were used as a training dataset 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 of <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 with men and women with 46,XY and 46,XX, respectively (figure 3A,B).

As a result, we detected 63 participants with discordance between self-reported sex and Affymetrix and/or PLINK sex calling (online supplemental table S2), then we examined their copy number variation (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 five participants with self-reported sex chromosome abnormalities including one mosaic Turner syndrome patient (45,X/46,XY) (scenarios 1 and 2). We identified all 48 participants with sex discordance as in the aforementioned sex check. For the 23 participants who had discordance

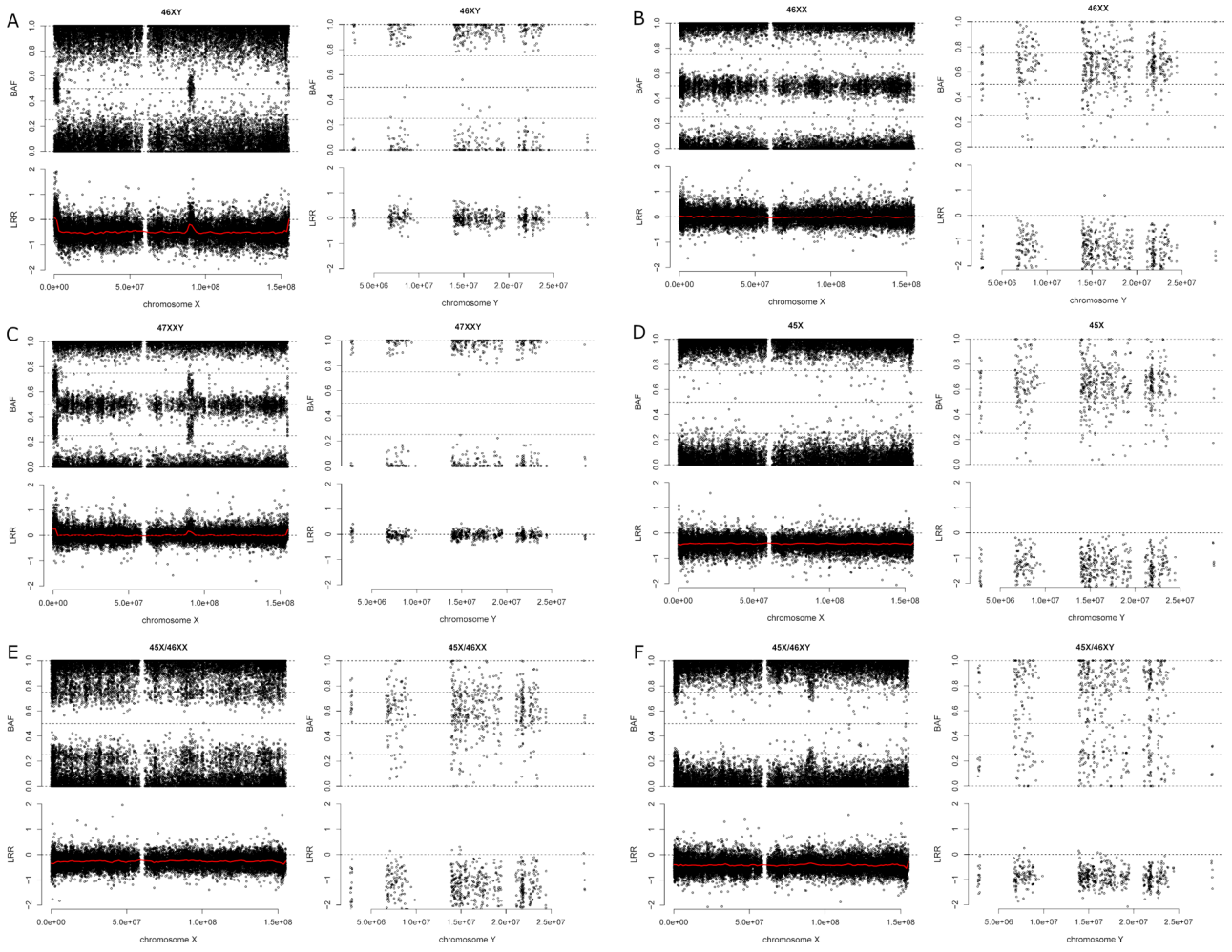

Figure 3 BAF (top) and log2 ratio (bottom) of chromosomes X and Y are shown for sex chromosome abnormalities. (A) In 46,XY, the BAF is either 0 or 1, and the expected log2 ratio is less than 0 on chromosome X. However, in the PAR and the chrY11.2/chrXq21.3 homology block, there are heterozygous calls in male shown as BAF of 0.5. The red line shows the locally weighted scatterplot smoothing 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 the log2 ratio is less than 0 on chromosome Y. (C) For Klinefelter syndrome (47,XXY), the log2 ratio is around 0 on chromosome X, which indicates ploidy as 2N. Compared with 46,XY, there are relatively lower peaks of log2 ratio at PAR and chrX21.3/chrY11.2 homology block region. Moreover, 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 a clear male pattern. (D) For Turner syndrome (45,X), on chromosome X, the log2 ratio is below 0, and there is no BAF bands of 0.5, which indicates one copy loss. Chromosome Y intensity profile showed a clear female pattern. (E) For 45,X/46,XX mosaicism, on chromosome X, there is a relatively smaller decrease of log2 ratio compared with one 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 a clear female pattern. (F) For 45,X/46,XY mosaicism, the log2 ratio is 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 indicate chromosome loss. However, the intermediate BAF values close to 0 or 1 at PAR and chrX21.3/chrY11.2 homology block region indicate the loss of chromosome Y existed in a larger proportion of cells. BAF, B allele frequency; PAR, pseudoautosomal region.

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 four participants with Klinefelter syndrome (47,XXY) and three with Turner Syndrome (45,X) (scenario 1) (figure 3C,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 3E,F). Additionally, individuals with low heterozygosity on chromosome X could be a result of inbreeding (online supplemental figure S7).

Charcot-Marie-Tooth (CMT) 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 nine CLSA participants who self-reported as having CMT. We examined their CNVs and found that four participants have duplication at *PMP22* (online supplemental figure S8), and one participant has deletion at *PMP22* (online supplemental figure S8). The other four subjects did not have CNVs detected at *PMP22*.

Human leukocyte antigen (HLA)-type imputation

We used the HLA*IMP:02 method[18] and a multipopulation reference panel[18] (ThermoFisher catalogue # 000.911) to impute HLA types. The genotypes of 11 MHC class I and class II loci with four-digit resolution were imputed for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4* and *HLA-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 with their known genotypes from literature. Calling accuracy was 100% across the loci (online supplemental table T3). 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 QC, and the set of 653 729 markers that passed all marker QC tests, with SNP-wise missingness of <0.05 and MAF of >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 V.r2, containing 97 256 reference samples at 308 107 085 genetic markers. We used this imputation service to prephase 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 three control samples. The two batches were 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 with the imputation using the Haplotype Reference Consortium reference panel containing 32 488 reference samples and 40.4 million genetic markers.[23] For each imputation dataset, 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 (online supplemental 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 prescreening 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 the 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://wwwclsa-elcvca/stay-informed/publications.

STRENGTHS AND LIMITATIONS

The CLSA genomic data are a unique resource nested in a large-scale, longitudinal study profiling the ageing 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 a certain threshold. In spite of these limitations, the CLSA cohort includes deep and extensive phenotyping and planned linkage to health administrative databases. For example, recently, the metabolomic data comprising 1314 biochemicals became available in approximately 9500 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 1479 participants. The CLSA has also initiated a subcohort to collect longitudinal data from MRI of the brain and microbiome of the gut in 6000 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 phenotypical 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 V.1.2 format using 8-bit encoding. It is recommended to use *qctool* V.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 leucocyte 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* (https://www.clsa-elcv.ca).

Author affiliations
[1]Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montréal, QC, Canada
[2]McGill Genome Centre, Department of Human Genetics, McGill University, Montréal, QC, Canada
[3]Department of Human Genetics, McGill University, Montréal, QC, Canada
[4]Hamilton Regional Laboratory Medicine Program, McMaster University, St. Joseph's Hospital St. Luke's Wing, Hamilton, ON, Canada
[5]Genetics & Genomic Biology, The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, ON, Canada
[6]Department of Medicine & of Epidemiology and Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada
[7]Montréal Heart Institute and Université de Montréal, Montréal, QC, Canada
[8]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
[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, NS, Canada
[11]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[12]Department of Bioengineering, McGill University, Montréal, QC, Canada

Twitter Lauren E Griffith @LaurenGriff1

Contributors VF and RL conducted data analyses and drafted the manuscript and share first authorship;CD-Z and AB generated data; JR supervised DNA extarction and genotyping data generation; CB, DR, CW, GL, GP, ADP, LEG, CV, ML, SK, PR, JBR and JR developed the concept and report design. All authors revised the manuscript critically for important intellectual content and approved the final version to be published. JR is the guarantor.

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

Competing interests None declared.

Patient consent for publication Not applicable.

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). The Canadian Longitudinal Study on Aging protocol was reviewed and approved by 13 research ethics boards across Canada. All participants provided written informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data are available from the Canadian Longitudinal Study on Aging (CLSA) (www.clsa-elcv.ca) for researchers who meet the criteria for access to deidentified CLSA data.

ORCID iDs
Rui Li http://orcid.org/0000-0001-8353-6772
Andrew D Paterson http://orcid.org/0000-0002-9169-118X
Lauren E Griffith http://orcid.org/0000-0002-2794-9692

REFERENCES

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;10:1249–52.
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;44:1090–7.
3 Singh PP, Demmitt BA, Nath RD, *et al*. The genetics of aging: a vertebrate perspective. *Cell* 2019;177:200–20.
4 Melzer D, Pilling LC, Ferrucci L. The genetics of human ageing. *Nat Rev Genet* 2020;21:88–101.
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:e1006977.
6 Raina P, Wolfson C, Kirkland S, *et al*. Cohort profile: the Canadian longitudinal study on aging (CLSA). *Int J Epidemiol* 2019;48:1752–3.
7 Affymetrix. UKB WCSGAX: UK Biobank 500K samples genotyping data generation by the Affymetrix research services laboratory, 2017. Available: 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;28:221–9.
9 Uk Biobank axiom array | UK Biobank. Available: http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ [Accessed 10 April 2018].
10 Manichaikul A, Mychaleckyj JC, Rich SS, *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26:2867–73.
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.
12 SPaC C. PLINK 1.9. Available: https://www.cog-genomics.org/plink1.9 [Accessed 27 April 2018].
13 , Auton A, Brooks LD, *et al*, 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;526:68–74.
14 Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005;76:887–93.
15 Galinsky KJ, Bhatia G, Loh P-R, *et al*. Fast Principal-Component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 2016;98:456–72.
16 Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7:781–91.
17 Price AL, Patterson NJ, Plenge RM, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–9.
18 Dilthey A, Leslie S, Moutsianas L, *et al*. Multi-population classical HLA type imputation. *PLoS Comput Biol* 2013;9:e1002877.
19 Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics* 2015;31:782–4.
20 Taliun D, Harris DN, Kessler MD, *et al*. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 2021;590:290–9.
21 Das S, Forer L, Schönherr S, *et al*. Next-Generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
22 Loh P-R, Danecek P, Palamara PF, *et al*. Reference-Based phasing using the haplotype reference Consortium panel. *Nat Genet* 2016;48:1443–8.
23 McCarthy S, Das S, Kretzschmar W, *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
24 GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease StudyBlindness GBD, Vision Impairment C. 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:e144–60.
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;108:1204–16.
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:e1003152.
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.
28 Campos AI, Ingold N, Huang Y. Genome-Wide analyses in 1,987,836 participants identify 39 genetic loci associated with sleep apnoea.
29 Uk Biobank axiom array, 2017. Available: 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;562:203–9.