The webinar, "Availability and quality assessment of genome-wide genetic data on 9,900 participants in the CLSA," will begin shortly.

For first-time WebEx users:

- This webinar will use VoIP only. Upon entering the session, you will be asked to join an integrated voice conference. Please select "yes". If you are not prompted with this message, please go to the main toolbar and select Audio>Integrated Audio Conference>Start Conference. If you continue to have audio issues, please go to Audio>Speaker/Microphone Audio Test.

- The only people in the session who can speak and be heard are the host and panelists.

- If you have questions/comments, you can type them into the chat box in the bottom right of the WebEx window. Ensure "All Participants" is selected from the dropdown menu before you press "send." Mobile users must select "Chat with Everyone." Questions will be visible to all attendees. You can type your questions at any point during the session, but they won't be answered until the end of the presentation.

- At the conclusion of the webinar, please consider participating in our feedback poll and remember to exit the WebEx session.

CLSA Webinar Series

Availability and quality assessment of genome-wide genetic data on 9,900 participants in the CLSA

Vince Forgetta, MSc, PhD

12 pm to 1 pm ET | January 15, 2019

Genetic and environmental factors contribute to maintaining health and in the development of disease and disability as people age. The Canadian Longitudinal Study on Aging (CLSA) presents a unique opportunity to study genetic and environmental contributions to human health and disease by providing information on the changing biological, medical, psychological, social, lifestyle and economic aspects of participants' lives. This webinar will describe the availability and quality assessment of genetic data for 9,900 CLSA participants, comprising genome-wide directly genotyped data for 794,409 markers and whole-genome imputed data for approximately 40 million genetic variants. Quality assessment of the genetic data include marker and sample-based tests, as well as analyses of sex-chromosome abnormalities, population structure, and familial relatedness.

As of fall 2018, qualified researchers from any country can access this genomic and phenotypic data release via the CLSA data access process.

Register online at: bit.ly/clsawebinars

Webinars will be broadcast using WebEx.
Further instructions will be sent by email.

clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

www.clsa-elcv.ca

Availability and quality assessment of genome-wide genetic data on 9,900 participants in the CLSA

January 15, 2018

The Team

Lady Davis Institute - McGill University

Vincenzo Forgetta, Brent Richards

McGill University and Genome Quebec Innovation Centre

Rui Li, Alexandre Belisle, Corinne Darmond-Zwaig, Jiannis Ragoussis, Mark Lathrop

Hamilton Regional Laboratory Medicine Program - McMaster University

Cynthia Balion, Guillaume Pare

The Hospital for Sick Children Research Institute

Delnaz Roshandel, Andrew D. Paterson

Epidemiology – McGill University

Christina Wolfson

Dalhousie University

Susan Kirkland

Montreal Heart Institute

Guillaume Lettre

Department of Health Research Methods, Evidence and Impact - McMaster University

Chris Verschoor, Parminder Raina

Outline

- Rationale
- Genetic Variation and Whole Genome Genotyping
- Genotyping of CLSA
- Whole Genome Imputation
- Genotype Imputation of CLSA
- Genome-wide Association Studies
- Example GWAS using CLSA
- Genomic Prediction
- Concluding Remarks

Rationale

2006

6 associations

5 studies

5 publications

GWAS Catalog (2018)

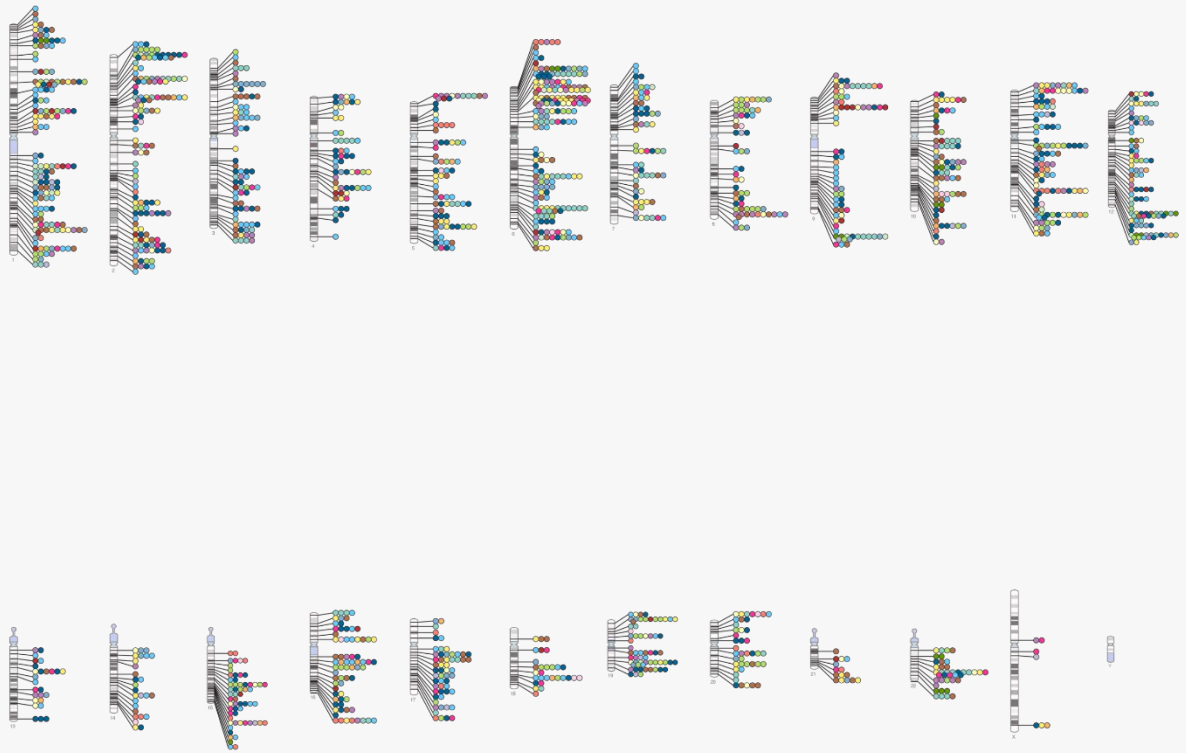
Rationale

2006 → 2010

4,677 associations

828 studies

718 publications



GWAS Catalog (2018)

Rationale

2006 → 2010 → 2018

69,000 associations

5,000 studies

3,378 publications

GWAS Catalog (2018)

Rationale

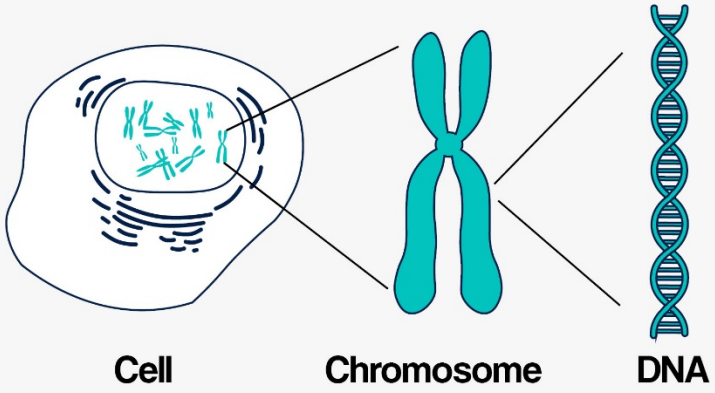
At 50,000 participants, the CLSA is amoung the largest and comprehensively assessed cohorts world-wide, providing an oppurtunity to investigate the genetics of health maintenance and development of disease and disability

Genome-wide Genotyping of CLSA

The Human Genome & Genetic Variation

The Human Genome

- Diploid: one chromosome per biological parent
- 22 pairs of autosomes, 2 sex chromosomes
- DNA: 3 billion nucleotides (As, Cs, Gs, and Ts)
- Encodes ~20,000 protein-coding genes



Human Genetic Variation

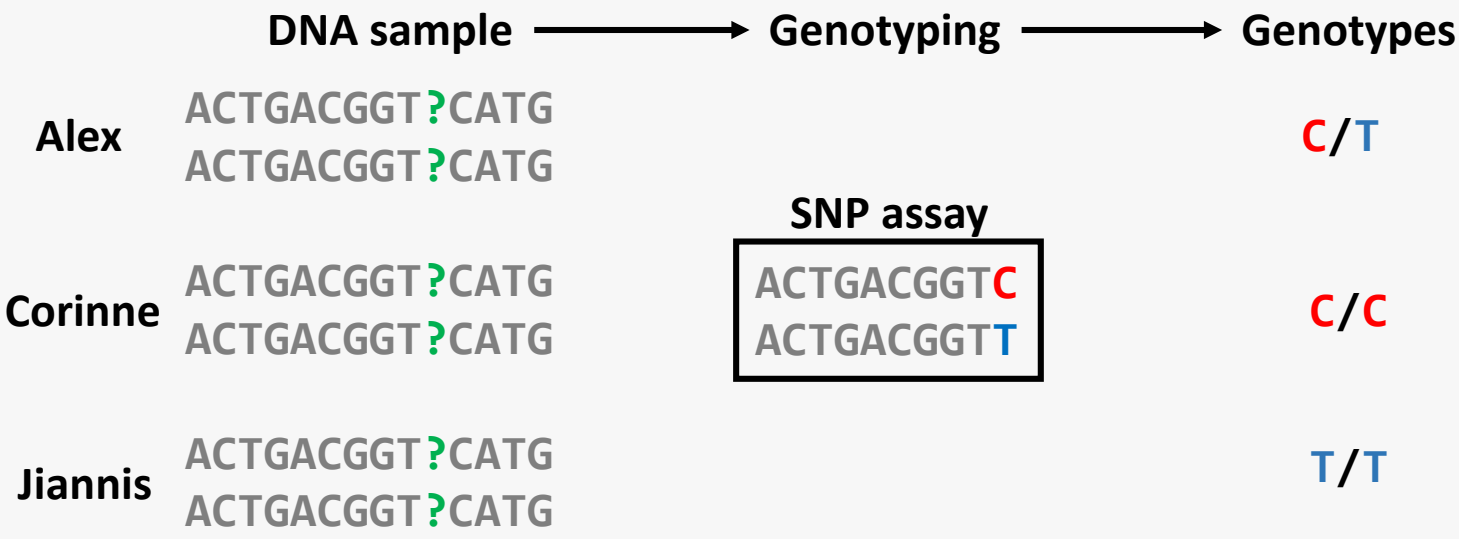
- Multiple types of variation: SNPs, indels, CNVs …
- SNP: Single nucleotide polymorphism
- >100 million SNPs in public databases

Vince	ACTGACGGTCCATG ACTGACGGTTCATG
Rui	ACTGACGGTCCATG ACTGACGGTCCATG
Brent	ACTGACGGTTCATG ACTGACGGTTCATG

Genotyping

How Genotyping Works

1. Given one or more DNA samples,
2. Determine sample genotype at a **known** SNP using pre-existing assay.
3. Assays can be multiplexed on “arrays” containing 500,000 to 1,00,000 SNPs.



Genome-wide Genotyping of CLSA

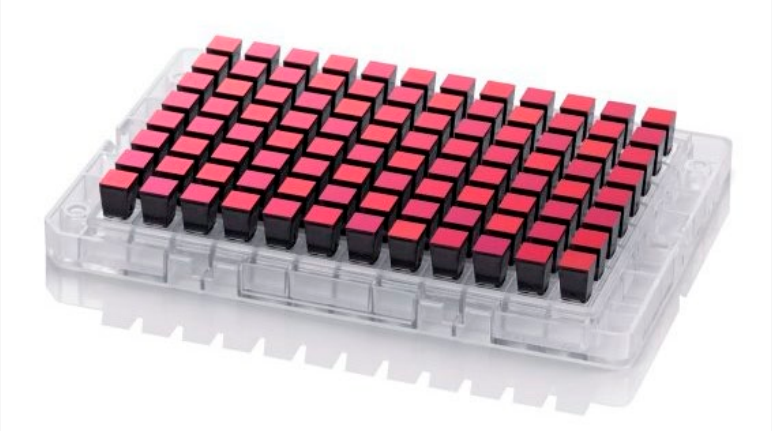
Affymetrix UK Biobank Axiom® Array

- Assay 820,967 SNP and indels
- Targets known disease associated SNPs
- Known protein-coding SNPs
- Panel of variants for optimal imputation in individuals of European ancestry

Genotyping of CLSA

- 9,900 individuals genotyped
- 2 batches of 5,092 and 5,048 individuals
- 794,409 SNPs per individual

96-sample assay plate



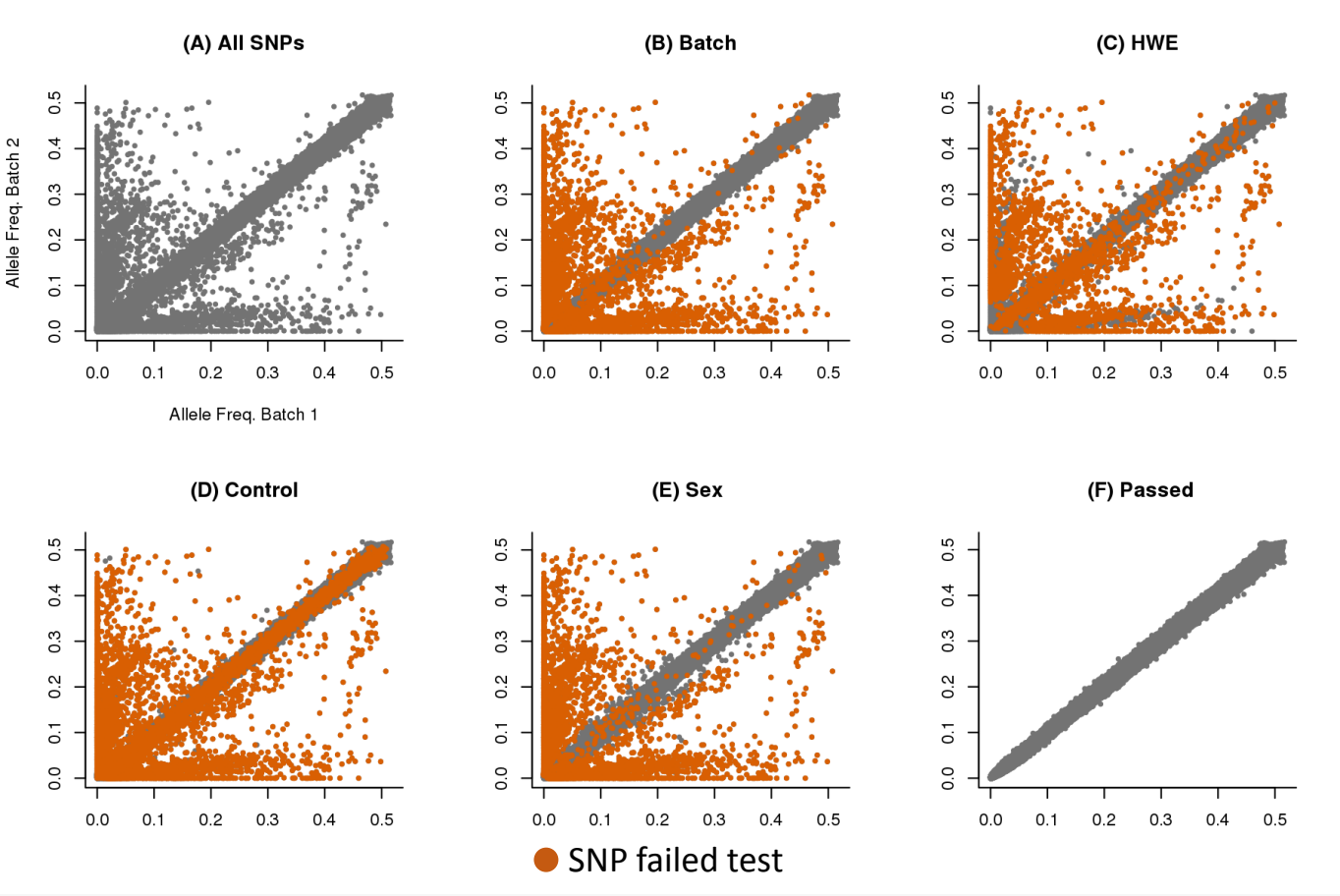
SNP Quality Assessment

Pre-Quality Assessment → **SNP-based tests** → **Post-Quality Assessment**

794,409 SNPs per individual

1. Batch-level genotype frequency
2. Hardy-Weinberg equilibrium
3. Control sample genotype frequency
4. Sex-wise genotype frequency

781,455 high quality SNPs per individual, 98.4% of total.

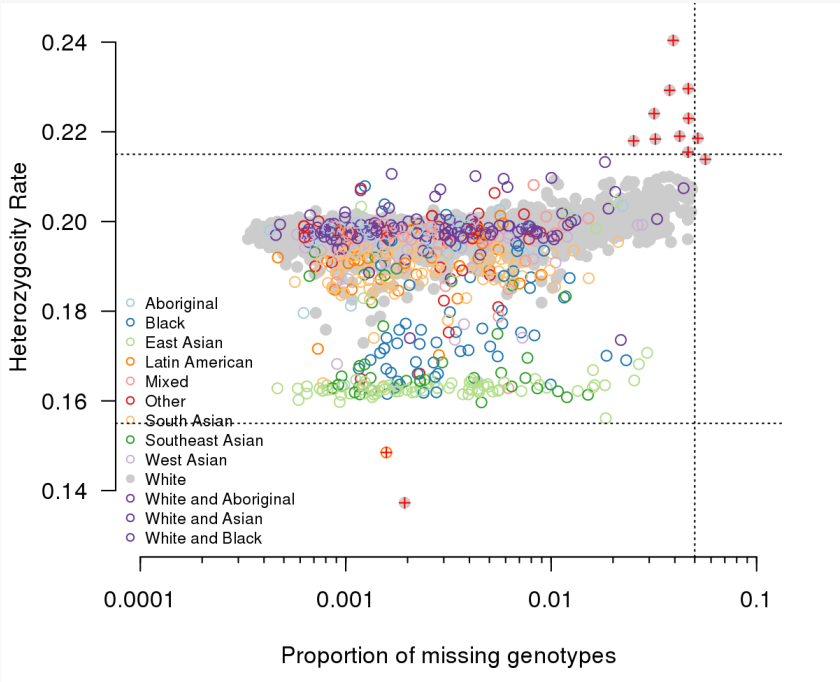


Sample Quality Assessment

Summary

1. Missingness: rate of missing calls per sample; elevated value indicates poor quality.
2. Heterozygosity: rate of heterozygous calls per sample; outliers indicates poor quality.
3. As expected, heterozygosity also associates with self-reported ancestry.

Heterozygosity versus missingness



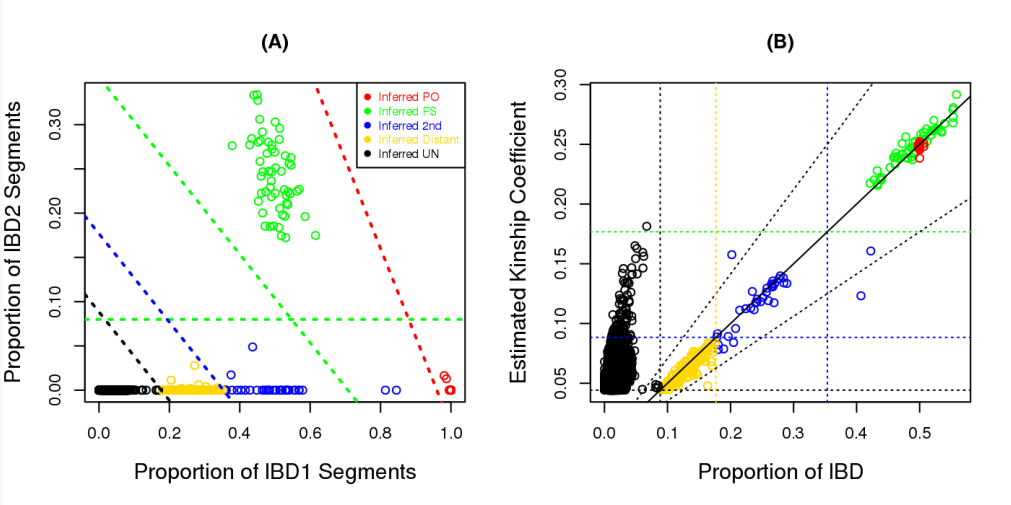
Other Population-level Analyses

Familial Relatedness

Summary

1. Not recorded by CLSA assessment, but is essential for some epidemiological and most genomic analyses.
2. Used KING software infer familial relatedness using genetic data.

Inferred Measures of Relatedness → **Counts of Inferred Relationships**



Inferred Relationship	Observed Pairs
2nd degree	35
3rd degres	137
Sibling	56
Parent-offspring	23
Unrelated	43643

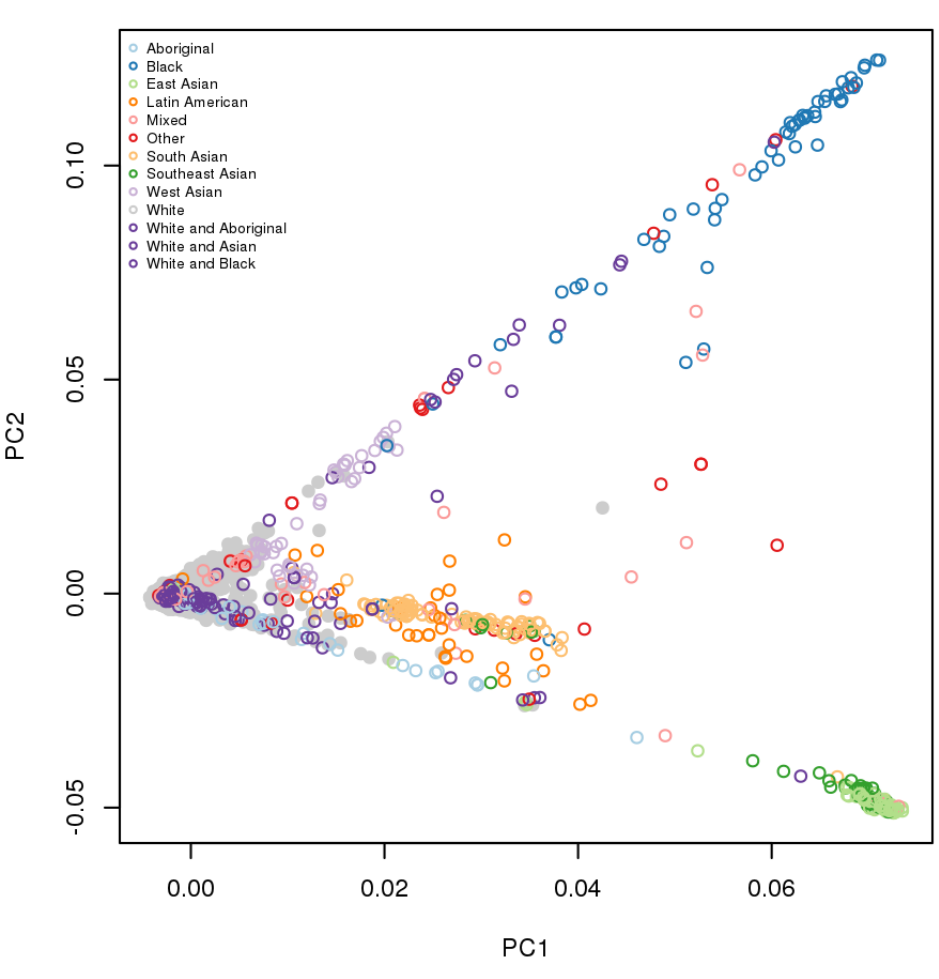
Population Structure

Purpose

To complement self-reported ancestry and used for controlling for population stratification in genome-wide association studies

Principal component analysis (PCA)

- PCA converts genotype data for all individuals into set of fewer components that capture the variability in the data.
- As expected, the top 2 PCs associate with self-reported ancestry.



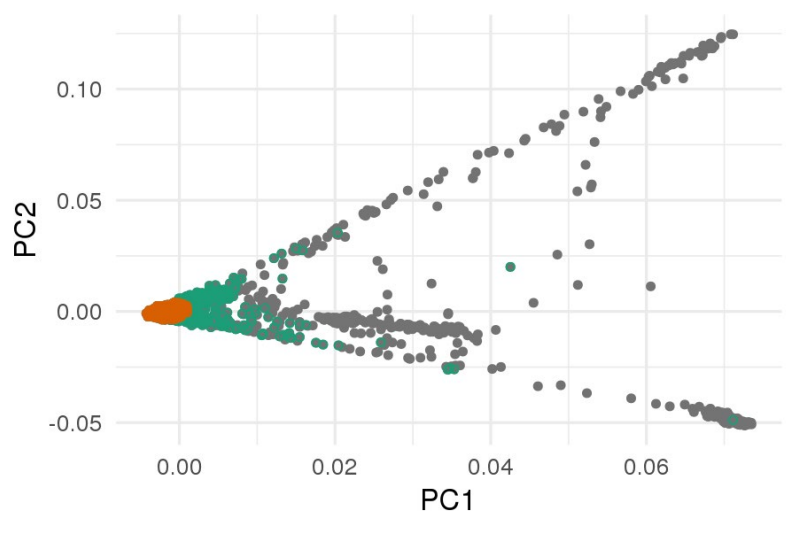
Largest ancestrally homogenous sub-population

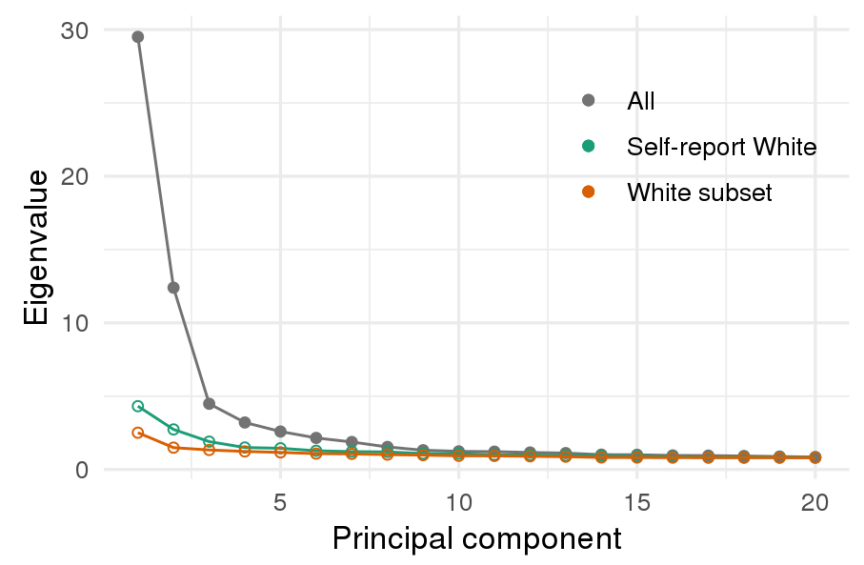
Purpose

To reduce the effect of population structure on downstream analyses it is recommended to use a subset of the population with relatively homogeneous ancestry.

Cluster analysis

- Cluster individuals using PCA components
- Dense cluster of individuals (N=8,209) overlaps self-reported White ancestry.



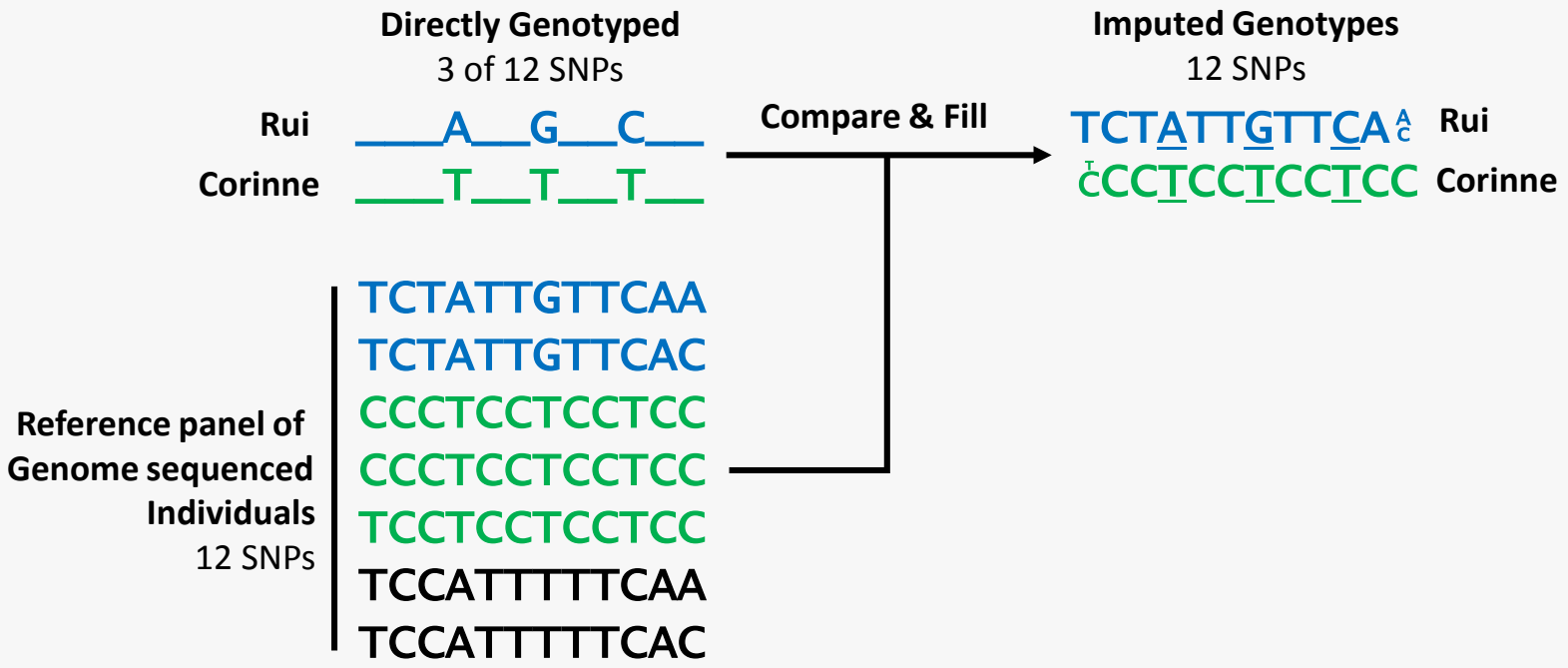


Genome-wide Imputation of CLSA

Whole-genome Imputation

Summary

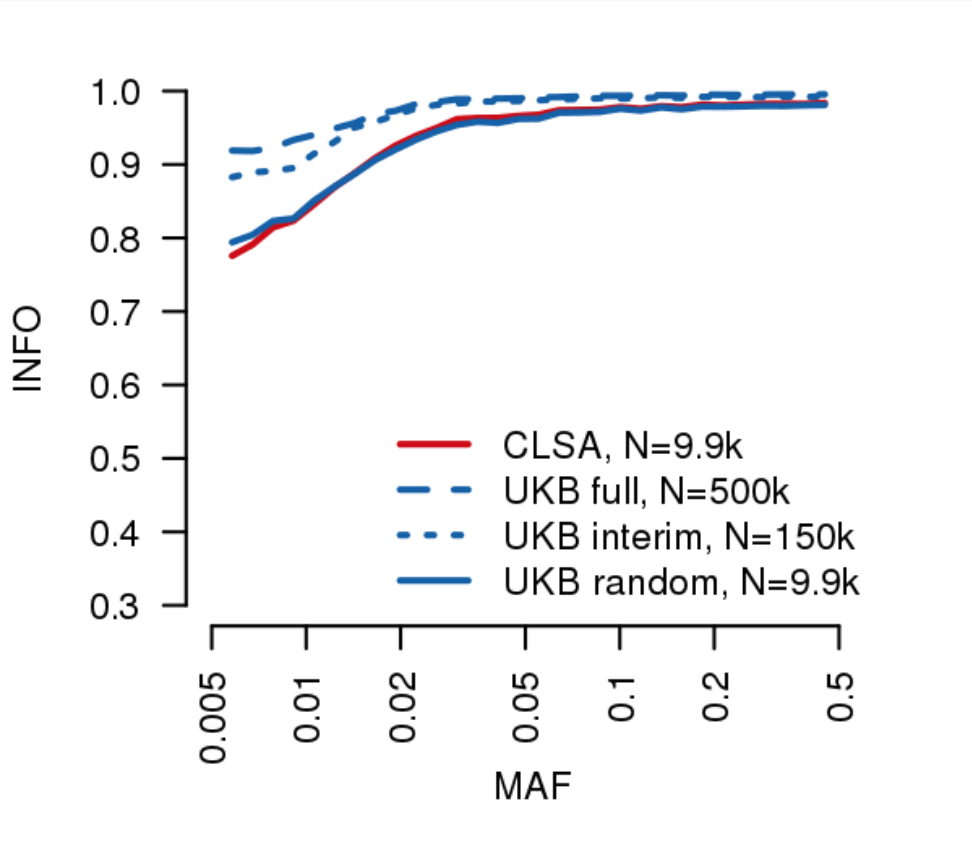
- To predict genotypes that are not directly assayed via the use of a densely genotyped reference panel.
- This denser set of imputed markers increases the power of genetic association studies and enables fine-mapping of causal variants.
- Imputation process produces 40-80 million genetic variants per individual.



Imputation of CLSA

Summary

- Using the Haplotype Reference Consortium reference panel containing ~32,000 individuals and 40.4 million SNPs.
- Impute 9,900 CLSA individuals using the online Sanger Imputation Service.
- **Result:** 40.4 million genetic variants for 9,900 participants, comparable in quality to UK Biobank.



Example Genome-wide Association Study

Genome-wide association study

Summary

- For each SNP, associate genotype to trait
- **Binary traits:** Increase in odds of disease per allele versus controls.
- **Continuous traits:** Unit increase in trait per allele.
- Methods typically adjust for population structure, familial relatedness, and other covariates.

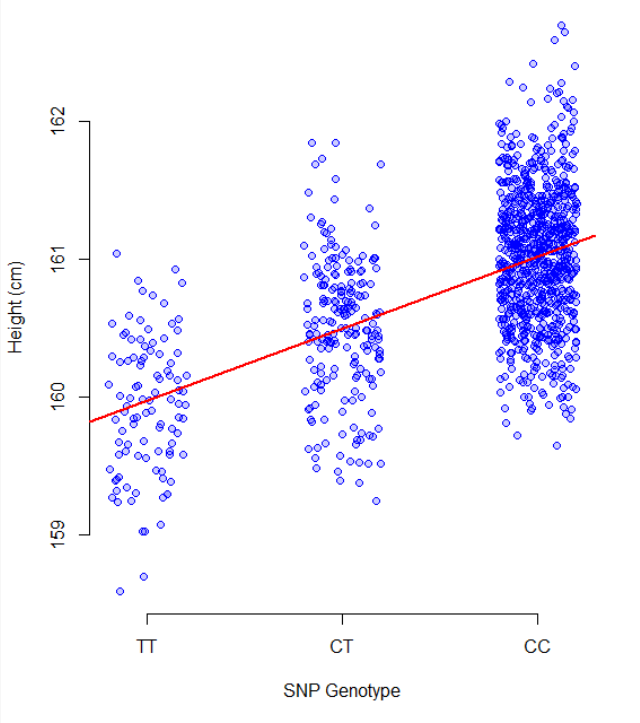
Binary Trait

	Case	Control
SNP allele T	a	b
SNP allele C	c	d

Odds Ratio = (a/b)/(c/d)

Relative Risk = (a/(a+b))/(c/(c+d))

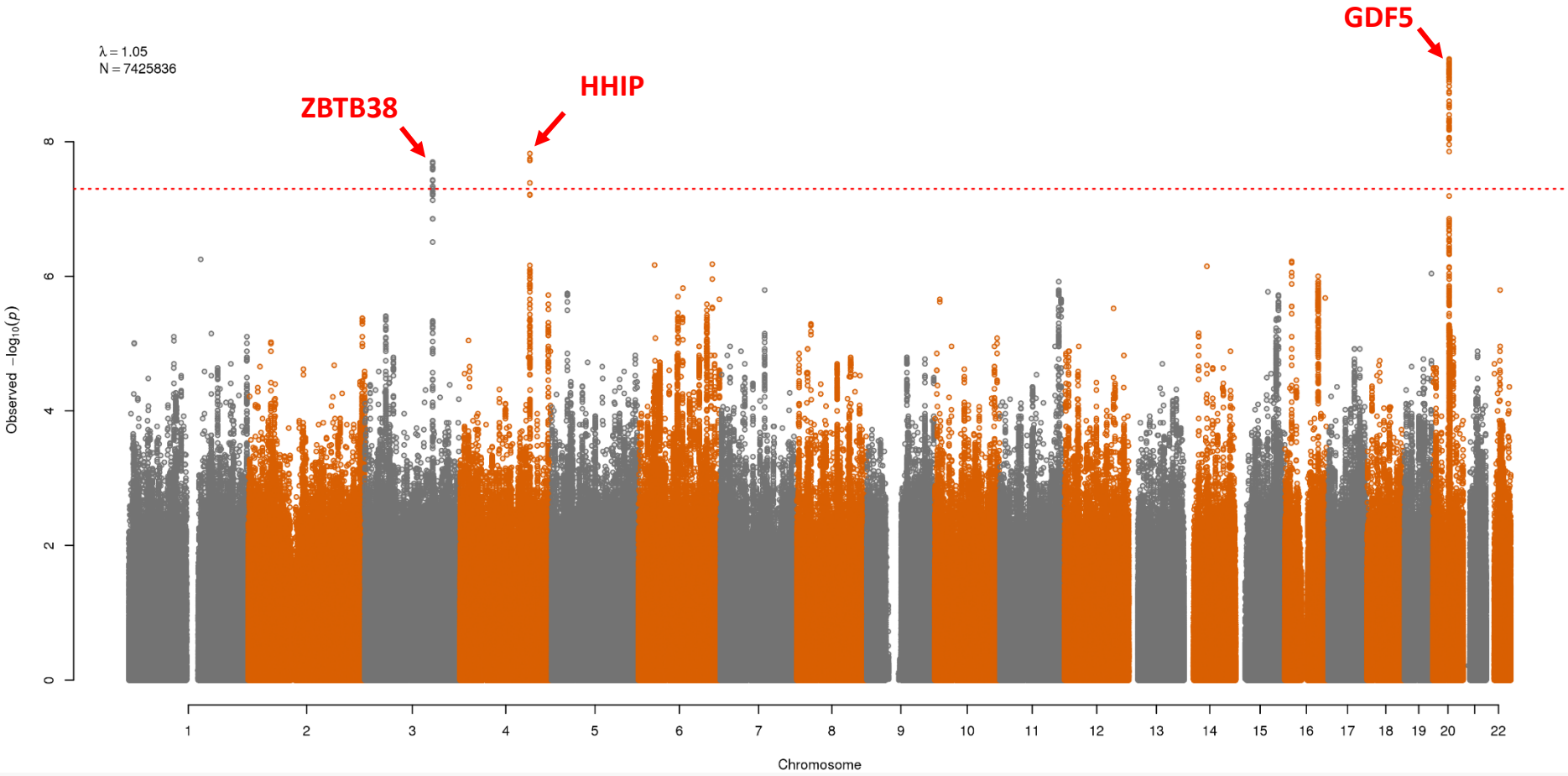
Continuous Trait



Height GWAS using CLSA

Summary

- 7.4 million association test statistics, one per SNP.
- Multiple-testing corrected p-value threshold is typically 1.2×10^{-8}.
- Compatible software: SNPTEST, BOLT-LMM, PLINK.



Genomic Prediction of Osteoporotic Fracture

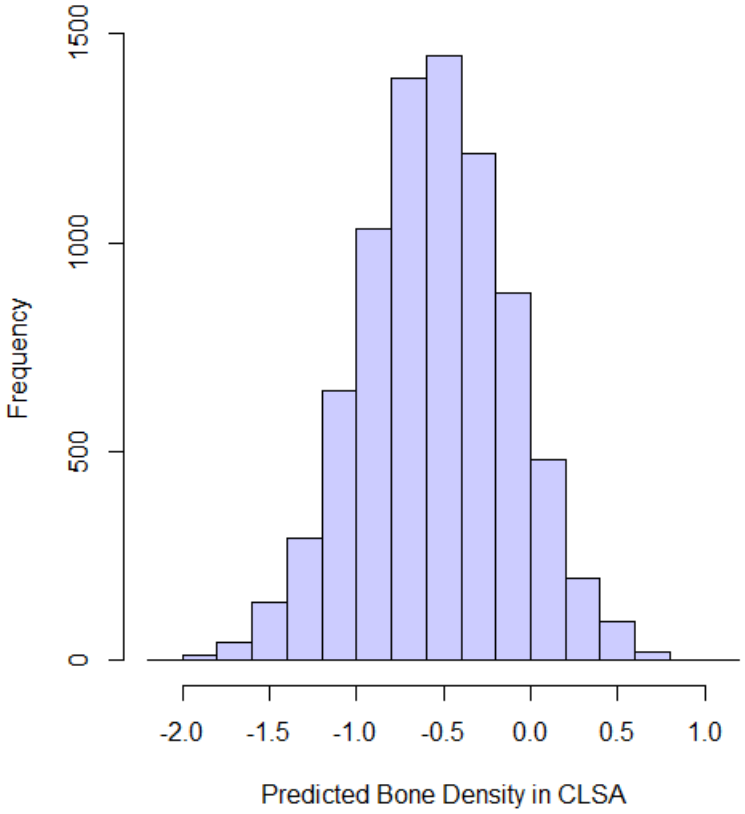
Genomic Prediction of Fracture

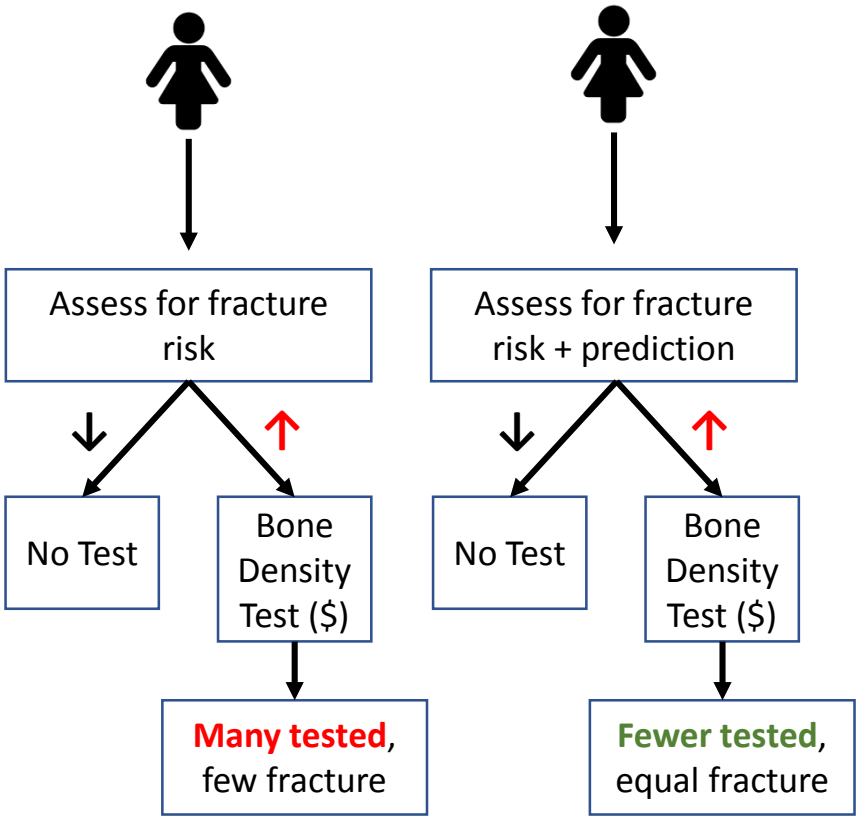
Summary

- We have developed a model to predict bone density using genetic data (Forgetta et al., 2018, bioRxiv).
- We show that it can predict fracture risk.
- We are using CLSA to validate results.

Why CLSA is Useful

- Large sample size, relevant population.
- Is assessed for all fracture risk factors.
- Measured bone density (gold standard).





Data formats and Software

Data Formats

Summary

- Data is made available using established genotype file formats.
- Binary format used to reduce download size and offers indexing.

Directly genotyped data

- 2 GB download
- Binary PLINK format
- **Software:** PLINK

Imputed Genotype data

- 36 GB download
- BGEN format (binary).
- **Software:** PLINK, BGENIX

	snp1	snp2	…	snpN
ind1	CC	TG	…	AT
ind2	GC	GG	…	AA
ind3	GG	TT	…	TT
ind4	GC	GG	…	AA
…	…	…	…	TA
indN	CC	GT	…	TT

	snp1			snp2		
	CC	CG	GG	GG	GT	TT
ind1	.97	.02	.01	.03	.95	.02
ind2	.03	.96	.01	.94	.03	.01
ind3	.03	.02	.95	.00	.02	.98
ind4	.01	.97	.02	.93	.05	.02
…	…			…		
indN	.99	.01	.00	.04	.92	.02

Concluding Remarks

Current release

- 9,900 individuals:
  - 794,409 directly genotyped SNPs.
  - 40.4 million imputed genotyped SNPs.
  - SNP-level and individual-level quality assessment statistics.
  - Imputation of HLA alleles.

Future release

- Additional 10,000 individuals.

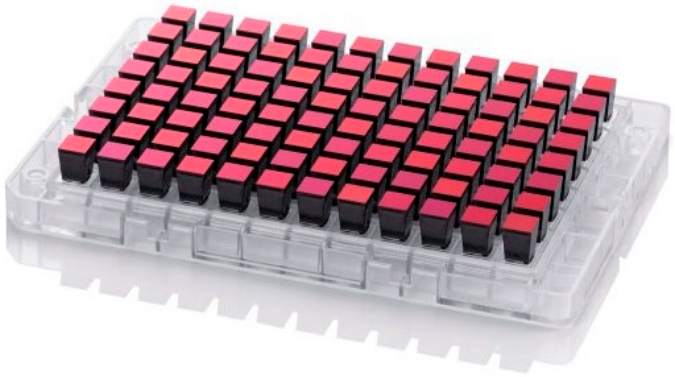
Questions?

Genome-wide Genotyping

Affymetrix UK Biobank Axiom® Array:

- Assay 820,967 SNP and indels
- Targets known disease associated SNPs
- Known protein-coding SNPs
- Panel of variants for optimal imputation in individuals of European ancestry

Assay for 96 samples



Upcoming CLSA Webinars

Characteristics of Caregivers and Care receivers in the Canadian Longitudinal Study on Aging

Debra Sheets, PhD

February 19, 2019 | 12 p.m. ET

Register: bit.ly/clsawebinars