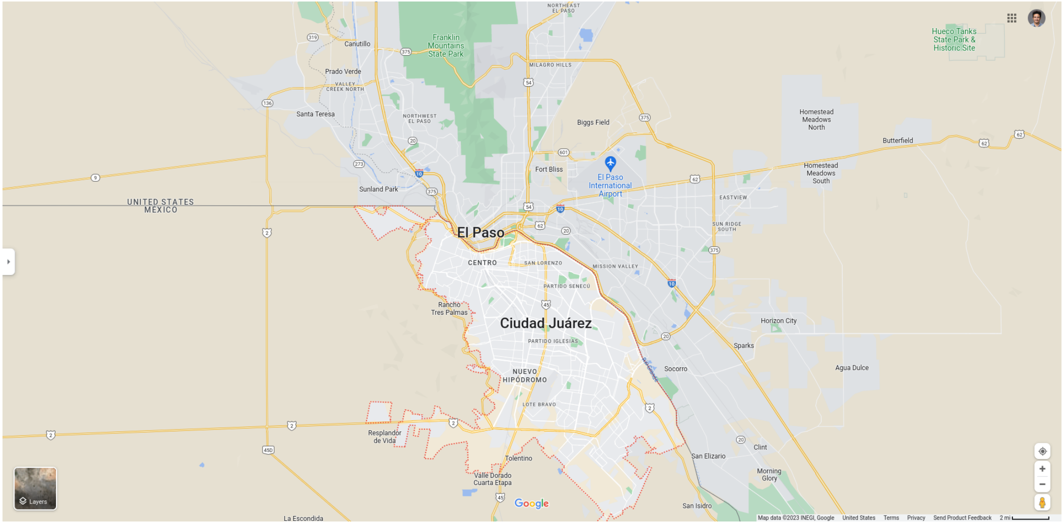
Genetic Association Studies and Population Structure in Nephrotic Syndrome

Alejandro Ochoa

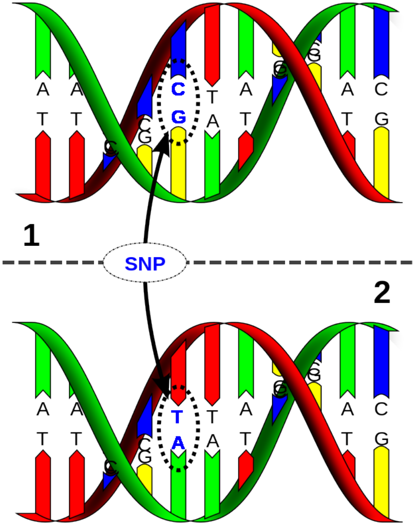
Biostatistics and Bioinformatics, StatGen — Duke University

2023-10-09

A little about me



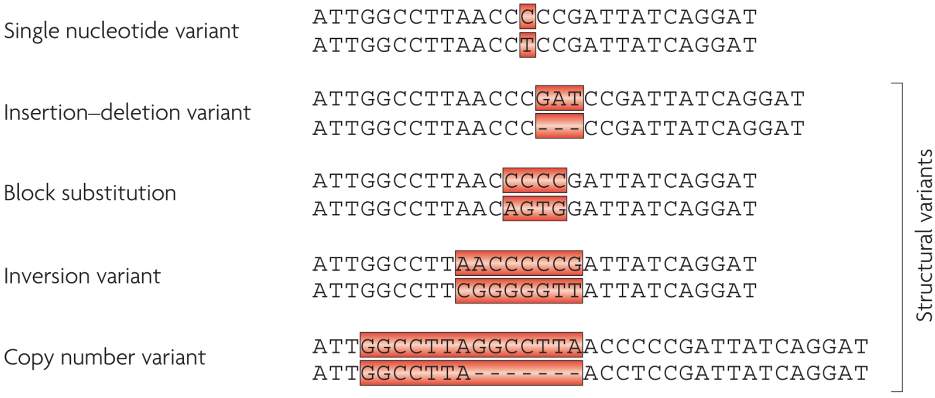
Genetic variation: we're all mutants!



Each newborn has ≈ 70 new mutations:

- Average mutation rate $\approx 1.1 \times 10^{-8}$ /base/generation
  - Higher in male lineage, with age
- Number of bases in genome $\approx 3.2 \times 10^{9}$, $\times 2$ for both copies

Types of mutations

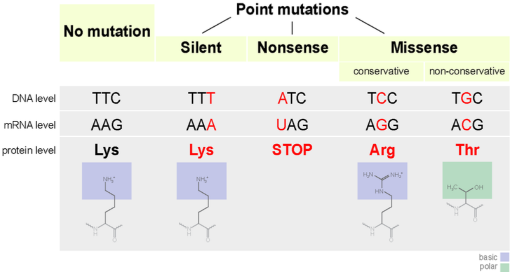

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Frazer *et al.* (2009)

- SNP = single nucleotide polymorphism
- Indel = insertion or deletion
- Structural variant = also large edits (gene or chr level)

Functional consequences of genetic variation

- Protein-coding mutation types



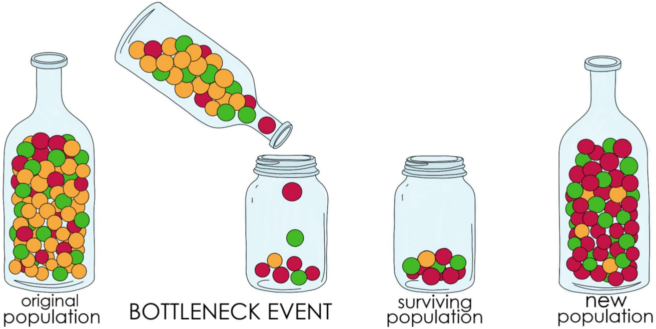
		Point mutations			
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr

Jonsta247, CC BY-SA 4.0, via Wikimedia Commons

- Non-coding mutations can affect gene expression

- Most are **neutral**:
  - Reveal relatedness and population history
- A small proportion cause disease
- Smallest proportion are beneficial:
  - New adaptation!

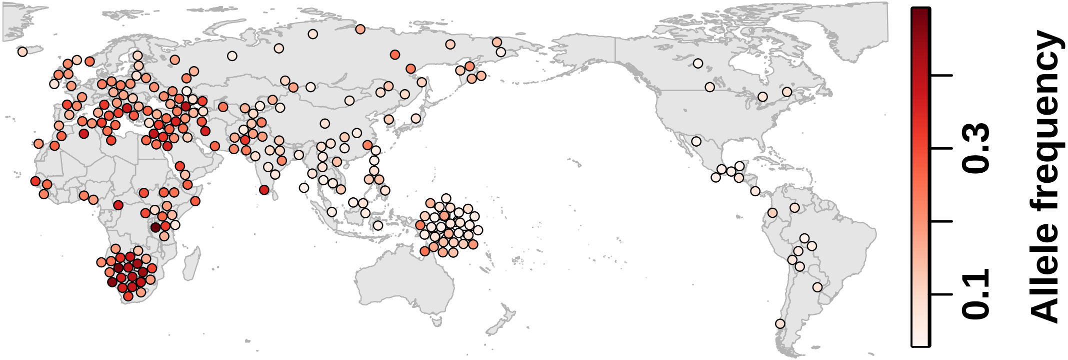
Dynamics of genetic variation



By Gabi Slizewska

- Most new mutations are lost
- Some become common in population
  - Outcomes are random
  - Variation greatest in small populations
  - Even disease alleles can become common

Human genetic structure: a typical SNP



Ochoa and Storey (2019a) doi:10.1101/653279

rs17110306; median differentiation given MAF $\geq$ 10%

Why? Migration and isolation, admixture, family structure

Every ancestry has genetic disease

- Disease variants are always arising spontaneously

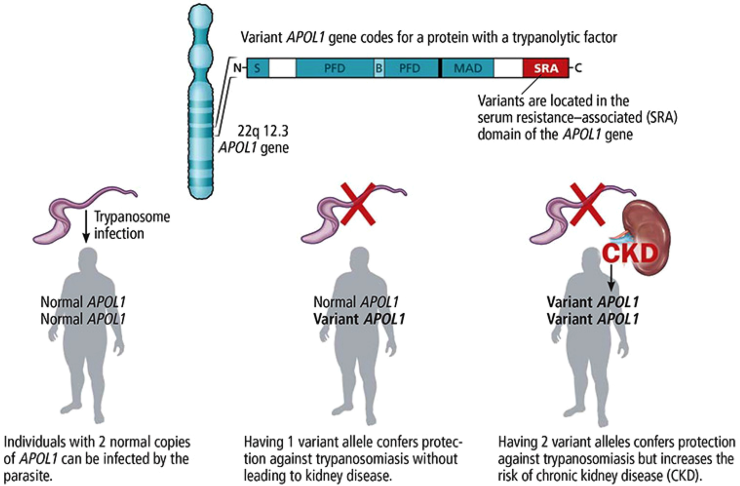
Every ancestry has genetic disease

- Disease variants are always arising spontaneously
- Selection gets rid of disease variants too slowly
  - Particularly for recessive and complex diseases

Every ancestry has genetic disease

- Disease variants are always arising spontaneously
- Selection gets rid of disease variants too slowly
  - Particularly for recessive and complex diseases
- Non-genetic causes of disease frequently also exist
  - “Environment”
  - Diet
  - Physical activity
  - Pollution
  - Racism
  - ...

APOL1 variants: beneficial heterozygotes, disease homozygotes

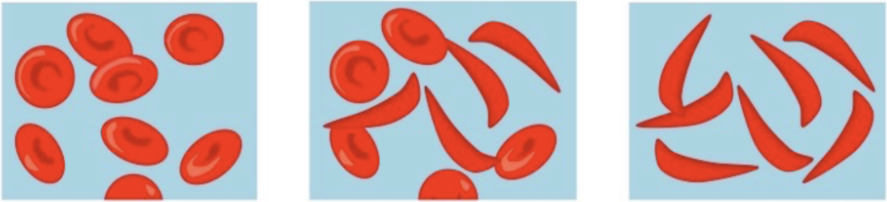


Variant *APOL1* gene codes for a protein with a trypanolytic factor
N
S
PFD
B
PFD
MAD
SRA
C
Variants are located in the serum resistance–associated (SRA) domain of the *APOL1* gene
22q 12.3
APOL1 gene
Trypanosome infection
CKD
Normal *APOL1*
Normal *APOL1*
Normal *APOL1*
Variant *APOL1*
Variant *APOL1*
Variant *APOL1*
Individuals with 2 normal copies of *APOL1* can be infected by the parasite.
Having 1 variant allele confers protection against trypanosomiasis without leading to kidney disease.
Having 2 variant alleles confers protection against trypanosomiasis but increases the risk of chronic kidney disease (CKD).

Variants in the APOL1 *gene that are common in sub-Saharan Africa protect against African sleeping sickness, but homozygosity for these variants increases the risk of CKD. Image taken with permission from J Nally Cleveland Clinic J of Medicine 2017*[47]

Smith and Brahman (2022)

Sickle cell disease: beneficial heterozygote, disease homozygote



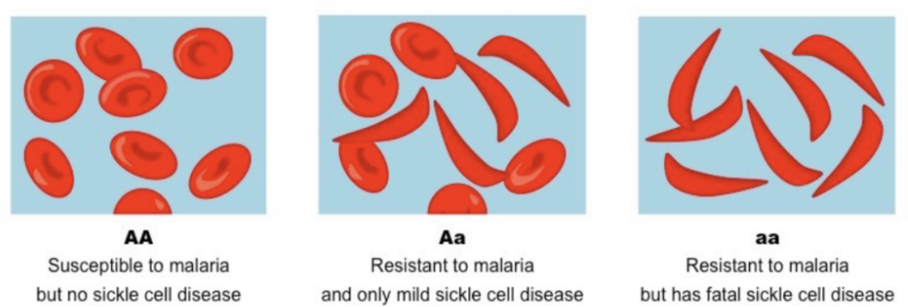
AA
Susceptible to malaria
but no sickle cell disease

Aa
Resistant to malaria
and only mild sickle cell disease

aa
Resistant to malaria
but has fatal sickle cell disease

chegg.com

Sickle cell disease: beneficial heterozygote, disease homozygote



AA
Susceptible to malaria
but no sickle cell disease

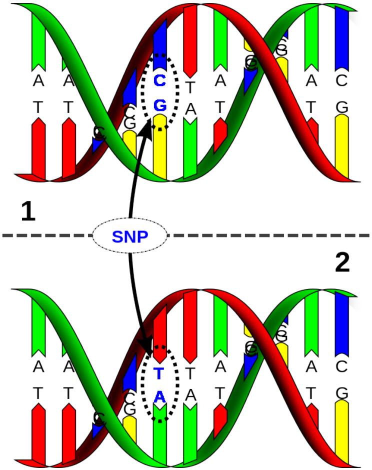
Aa
Resistant to malaria
and only mild sickle cell disease

aa
Resistant to malaria
but has fatal sickle cell disease

chegg.com

Additional variants in BCL11A and elsewhere can ameliorate SCD!

Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
CC	0
CT	1
TT	2

⇒

Individuals

Loci

0 2 2 1 1 0 1
0 2 1 0 1
2 …

$\mathbf{X}$

Hardy-Weinberg Equilibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

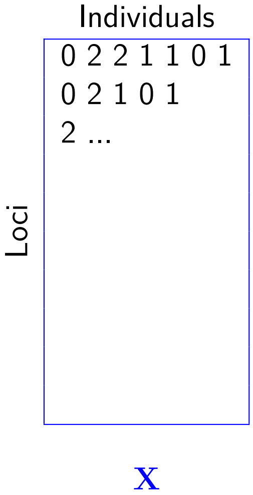
p_i = frequency of reference allele at locus i.

Under HWE:

$$\begin{aligned}
\Pr(x_{ij} = 2) &= p_i^2, \\
\Pr(x_{ij} = 1) &= 2p_i \left(1 - p_i\right), \\
\Pr(x_{ij} = 0) &= \left(1 - p_i\right)^2.
\end{aligned}$$

HWE not valid under genetic structure!

Dependence structure of genotype matrix



0 2 2 1 1 0 1
0 2 1 0 1
2 …

High-dimensional binomial data

- No general likelihood function
- My work: method of moments

Relatedness / Population structure

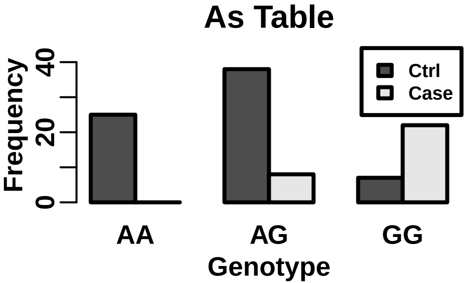
- Dependence between individuals (columns)

Linkage disequilibrium

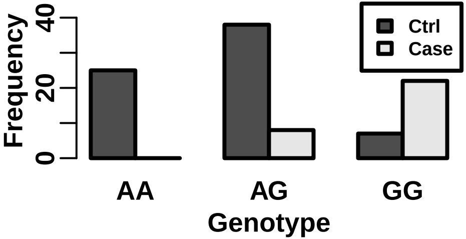
- Dependence between loci (rows)

Genetic association study: genotype-phenotype correlation

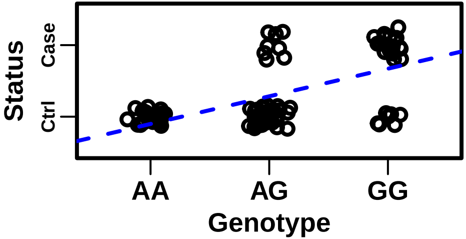
Genetic association study: genotype-phenotype correlation



Genetic association study: genotype-phenotype correlation



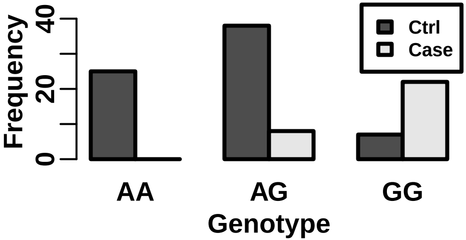
As Table



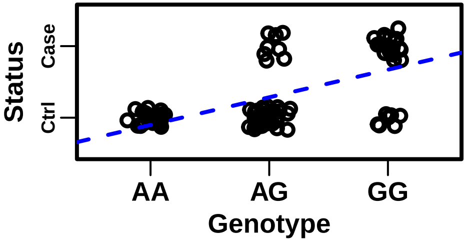
As Regression

Genetic association study: genotype-phenotype correlation

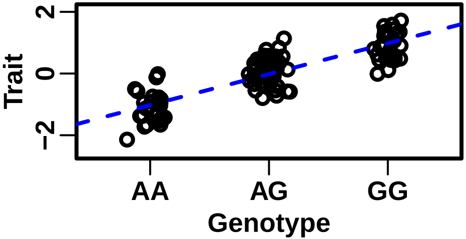
As Table



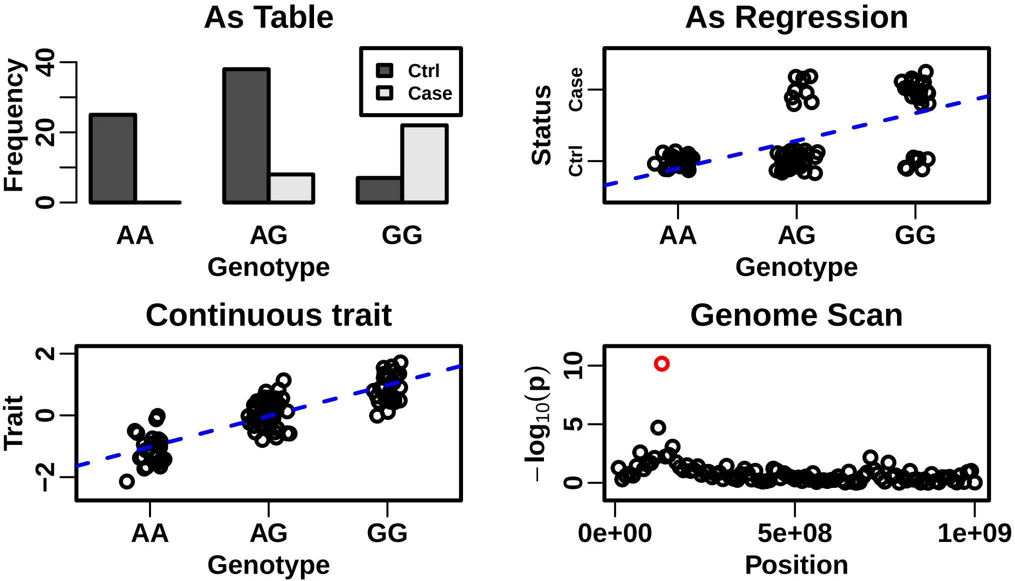
As Regression



Continuous trait



Genetic association study: genotype-phenotype correlation



As Table

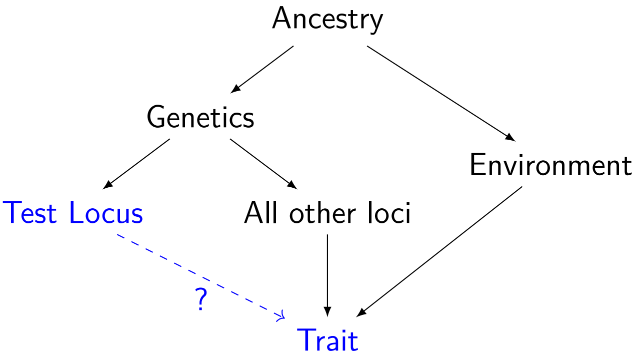
As Regression

Continuous trait

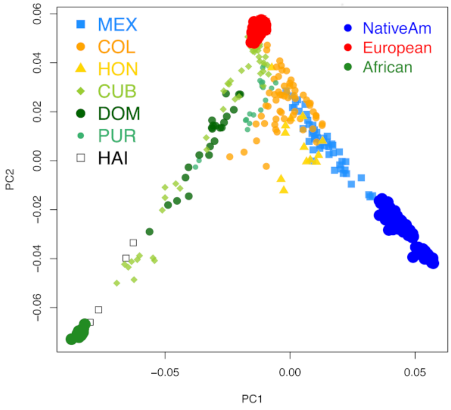
Genome Scan

Why is this problem so hard?

- Millions of tests
- Polygenicity (many causal variants)
- Confounders
- Incorrect assumptions: independence / additivity



PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

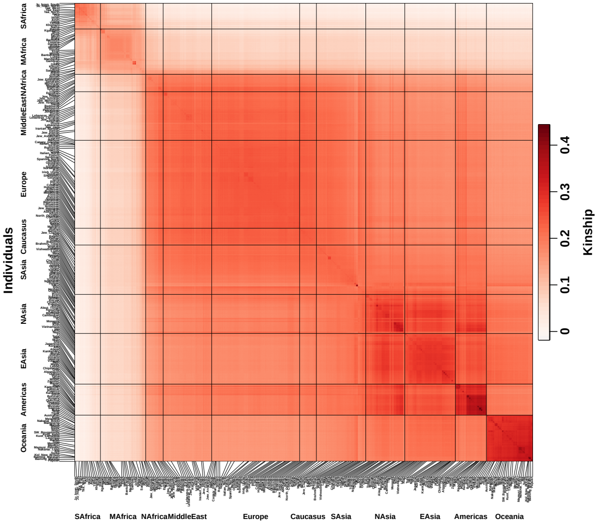
PCs map to ancestry.

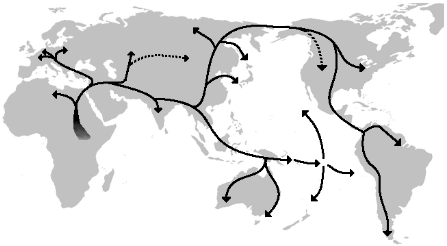
”PCs” are top eigenvectors of kinship matrix.

Pros: Fast!

Cons: Fails on family data.

Kinship (covariance) matrix of world-wide human population





Ochoa and Storey (2019) doi:10.1101/653279

Association with PCA vs LMM

Principal Components Analysis (PCA)
and Linear Mixed-effects Model (LMM):

$$\begin{aligned}
\text{PCA}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon, \\
\text{LMM}: \quad & \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.
\end{aligned}$$

$\mathbf{U}_d$ are top d eigenvectors of kinship matrix Φ.
$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\Phi\right)$.

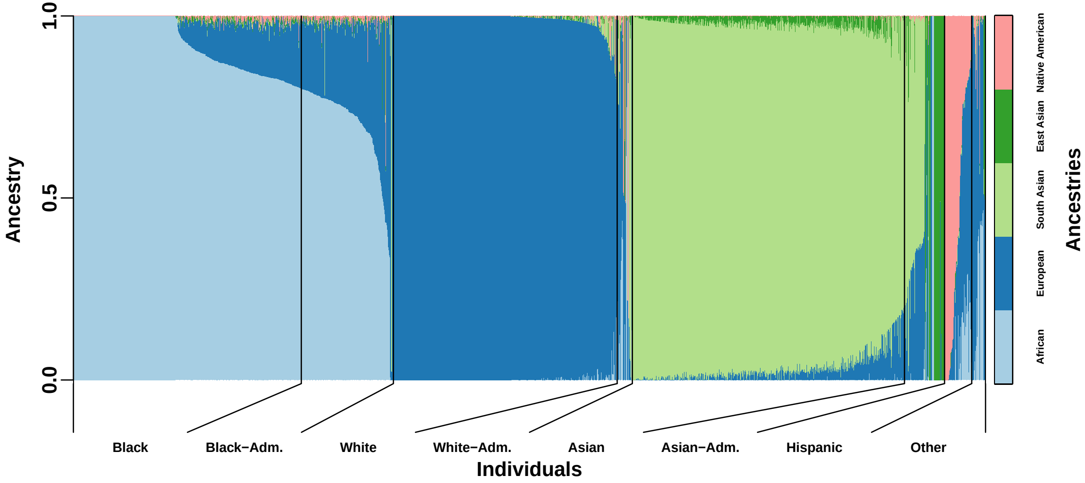
- PCA is faster but low-dimensional
- LMM is slower but can model families

Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

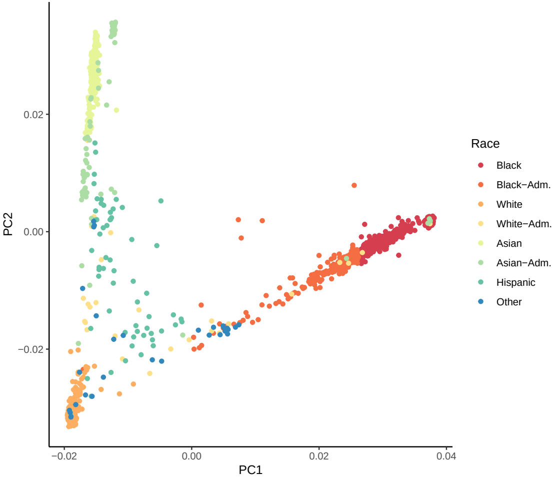
Nephrotic Syndrome association study

- Severe pediatric kidney disease.
- 1,000 cases/1,000 controls
- Multiethnic
  - Diverse Duke patients
  - Nigeria
  - Sri Lanka
- Included all 2,504 samples from 1000 Genomes as additional controls

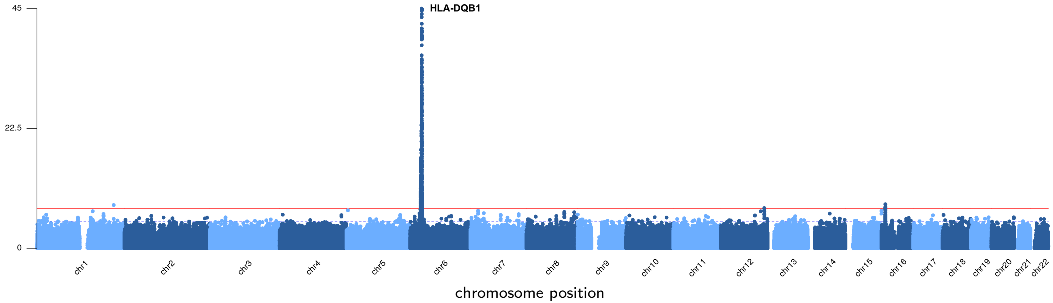
Nephrotic Syndrome association study: Admixture plot



Nephrotic Syndrome association study: PCA plot



Nephrotic Syndrome association study: Manhattan plot


HLA-DQB1
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chromosome position