Genetic Association Studies and Population Structure in Nephrotic Syndrome

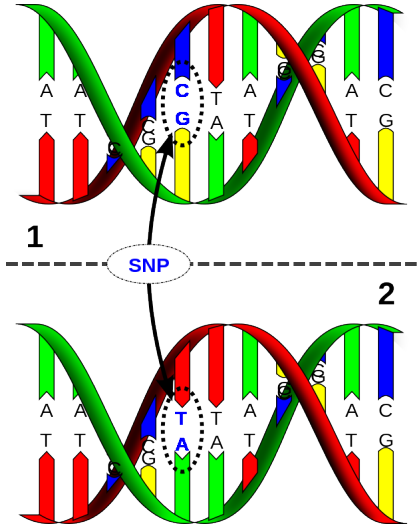
Alejandro Ochoa

Duke Center for Statistical Genetics and Genomics

Department of Biostatistics & Bioinformatics
Duke University School of Medicine

2025-01-30 — Genomics Resource Workshop

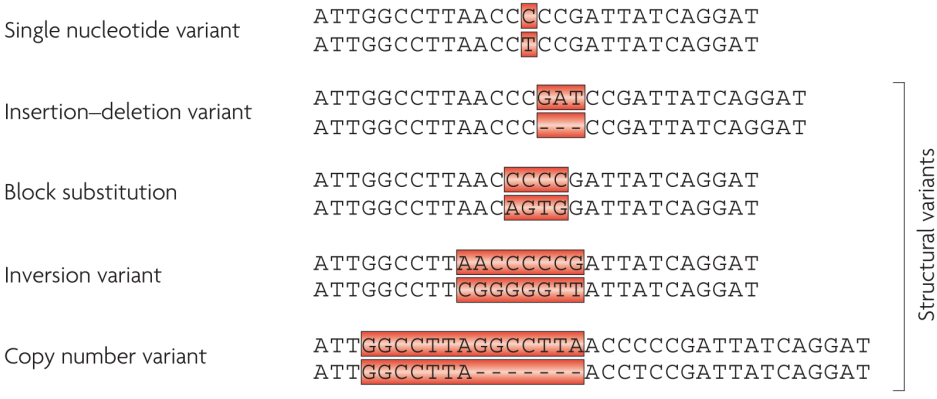
Genetic variation: we're all mutants!



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

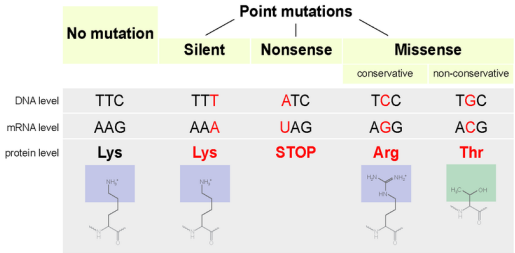


Frazer *et al.* (2009)

- 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

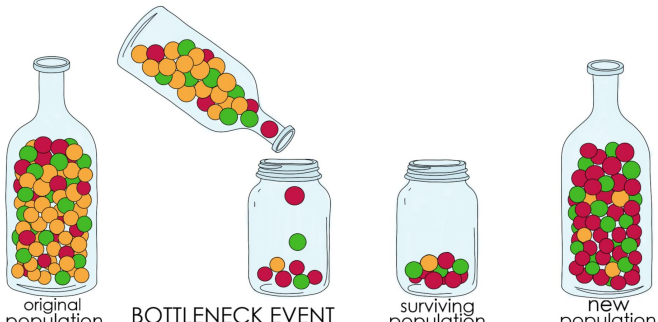


Jonsta247, CC BY-SA 4.0, via Wikimedia Commons

- 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!

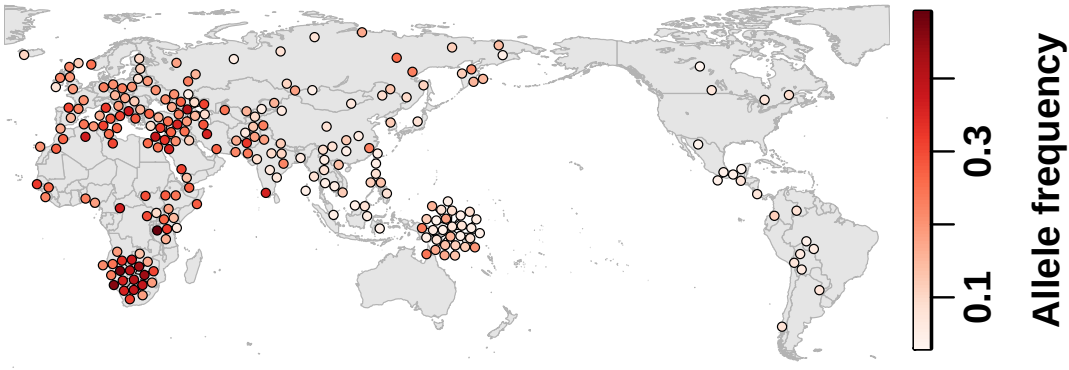
Dynamics of genetic variation



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

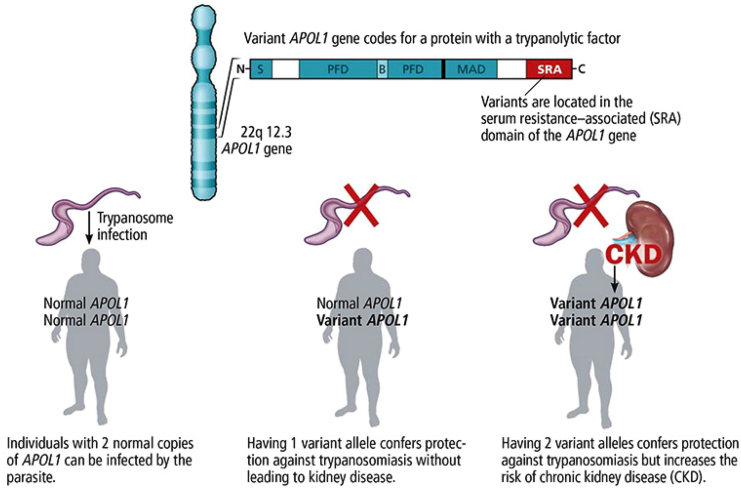
Every ancestry has genetic disease

- 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
  - ...

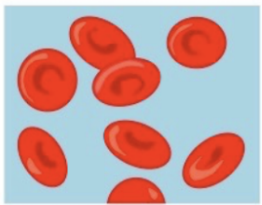
APOL1 variants: beneficial heterozygotes, disease homozygotes



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)

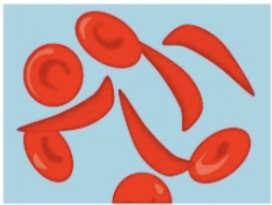
Sickle cell disease: beneficial heterozygote, disease homozygote



AA

Susceptible to malaria

but no sickle cell disease



Aa

Resistant to malaria

and only mild sickle cell disease

aa

Resistant to malaria

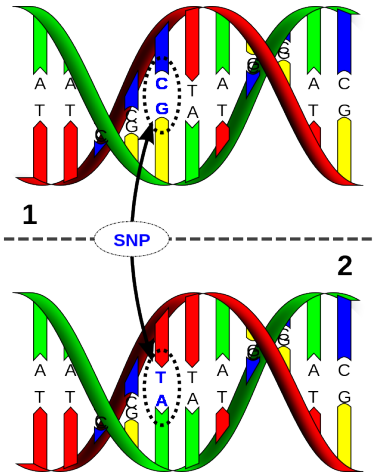
but has fatal sickle cell disease

chegg.com

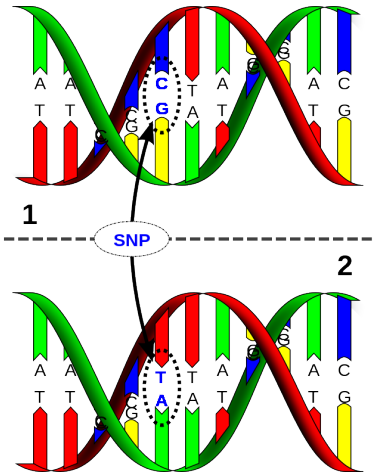
Sickle cell disease: beneficial heterozygote, disease homozygote

Additional variants in BCL11A and elsewhere can ameliorate SCD!

Single Nucleotide Polymorphism (SNP) data



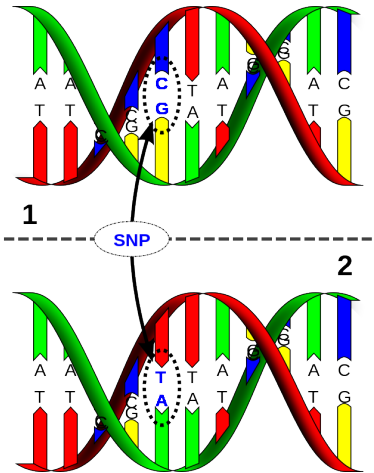
Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

$\Rightarrow$

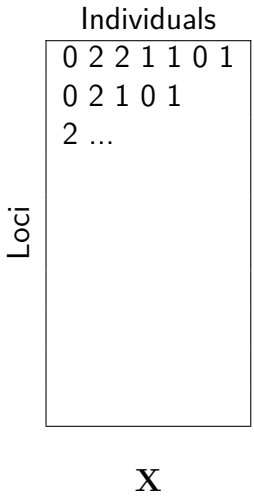
Individuals

Loci

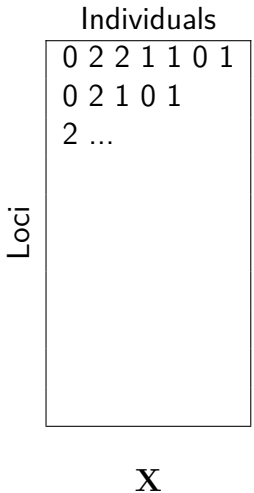
0 2 2 1 1 0 1
0 2 1 0 1
2 ...

$\mathbf{X}$

Dependence structure of genotype matrix



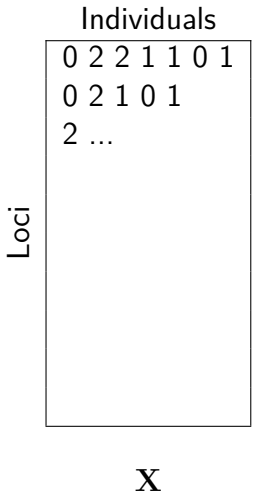
Dependence structure of genotype matrix



Relatedness / Population structure

- Dependence between individuals (columns)

Dependence structure of genotype matrix



Relatedness / Population structure

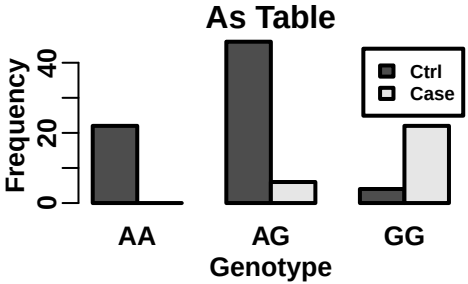
- Dependence between individuals (columns)

Linkage disequilibrium

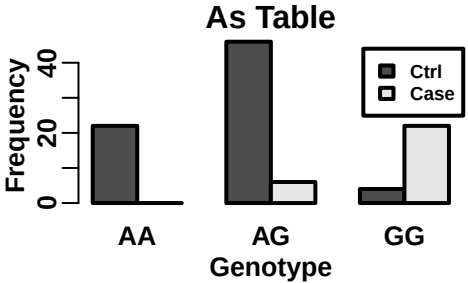
- Dependence between loci (rows)

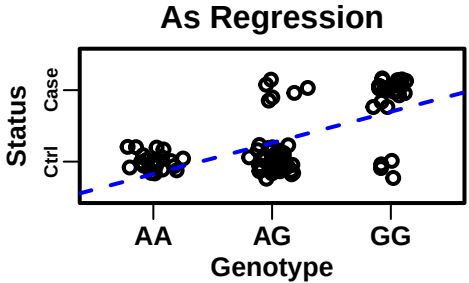
Genetic association study: genotype-phenotype correlation

Genetic association study: genotype-phenotype correlation

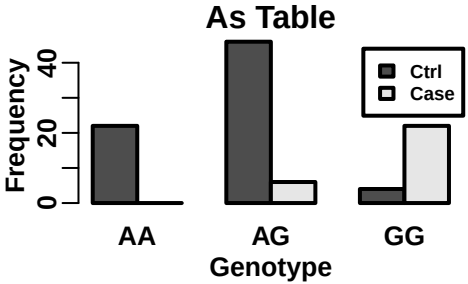


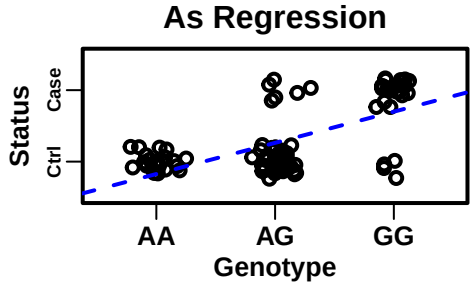
Genetic association study: genotype-phenotype correlation

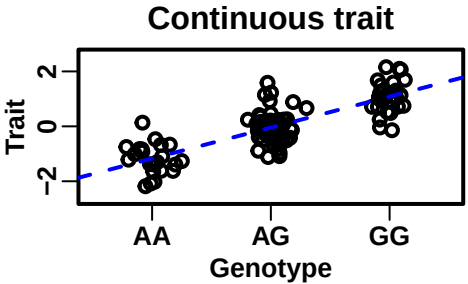




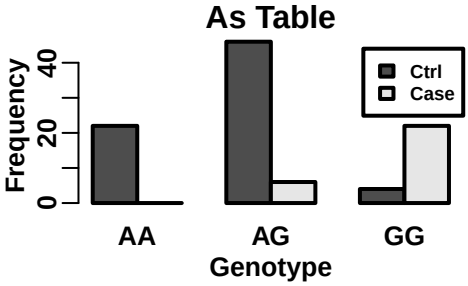
Genetic association study: genotype-phenotype correlation

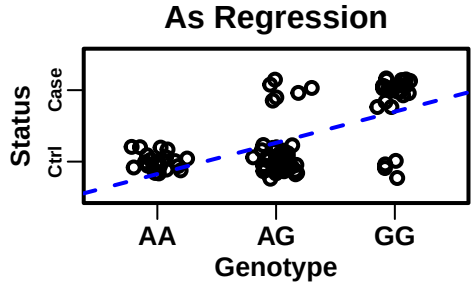


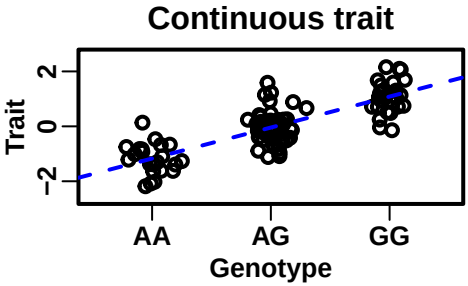


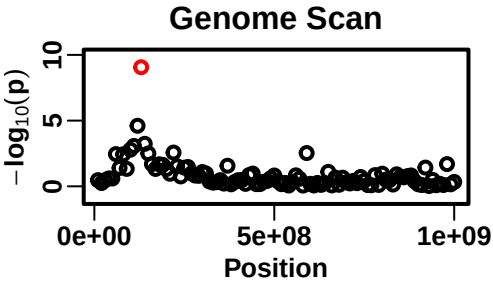


Genetic association study: genotype-phenotype correlation









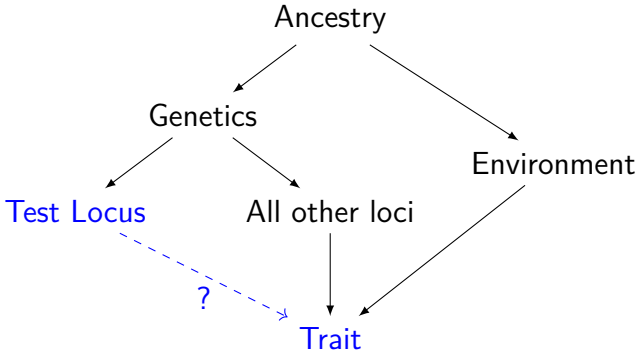
Why is this problem so hard?

Why is this problem so hard?

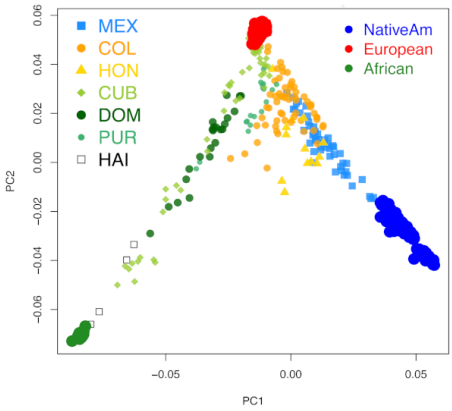
- Millions of tests
- Polygenicity (many causal variants)
- Incorrect assumptions: independence / additivity
- Confounders

Why is this problem so hard?

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



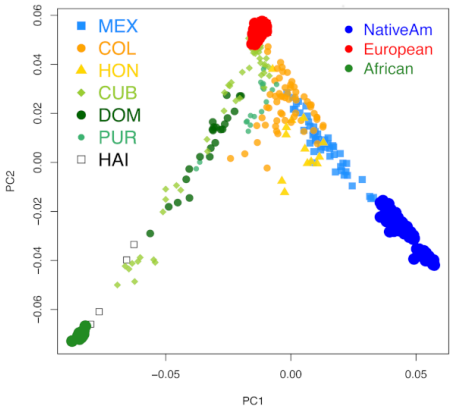
PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

PCA: Principal Component Analysis

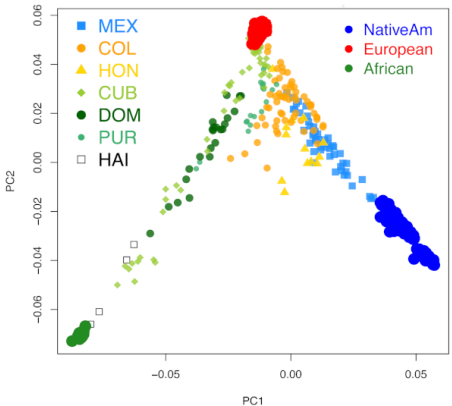


Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

PCs map to ancestry.

PCA: Principal Component Analysis



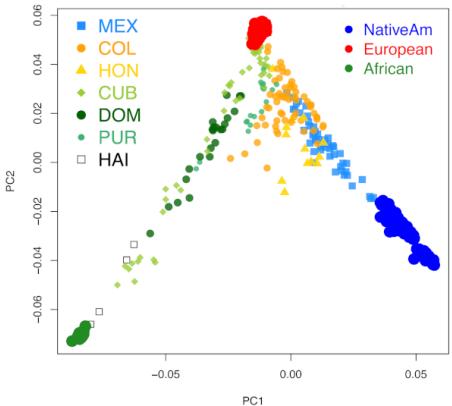
Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

PCs map to ancestry.

"PCs" are top eigenvectors of kinship matrix.

PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

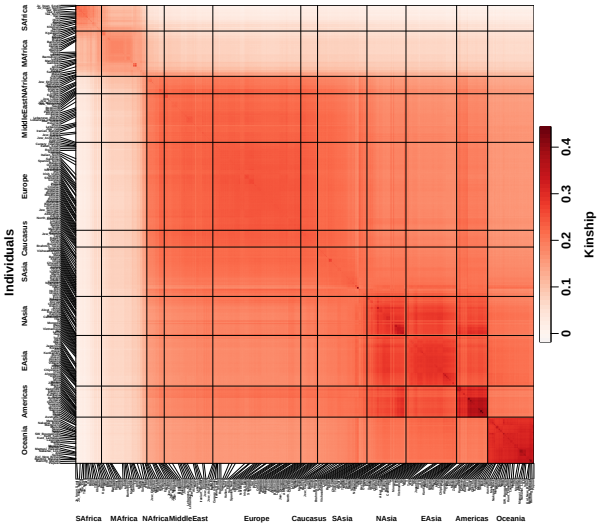
PCs map to ancestry.

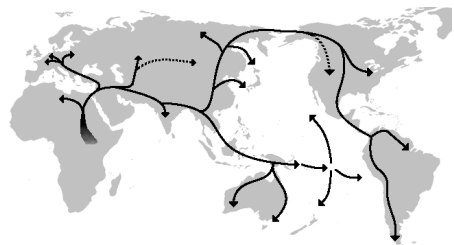
"PCs" are top eigenvectors of kinship matrix.

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):

Association with PCA vs LMM

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

$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$

$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

Association with PCA vs LMM

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

$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$

$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

$\mathbf{U}_d$ are top d eigenvectors of kinship matrix Φ.

$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\Phi\right)$.

Association with PCA vs LMM

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

$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$

$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

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

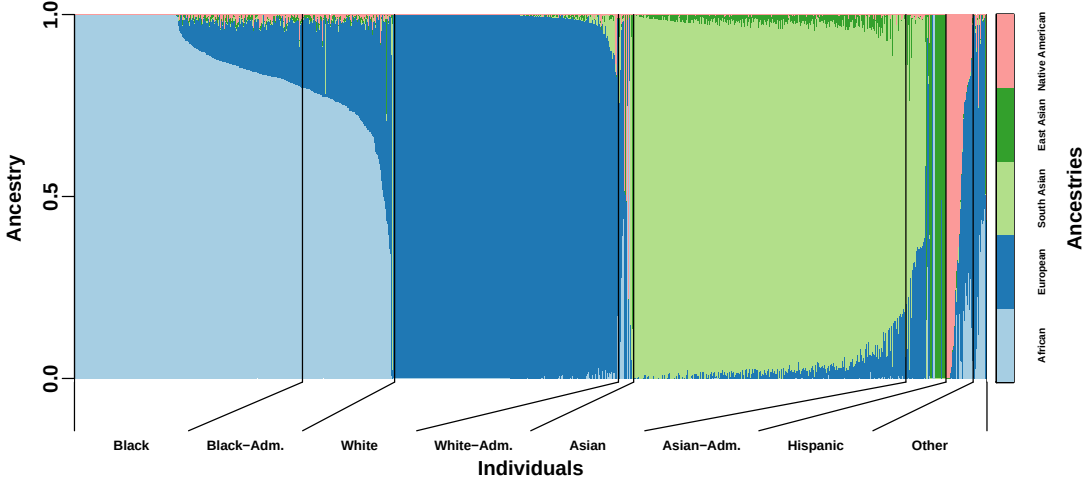
- PCA is faster but low-dimensional
- LMM is slower but can model families

Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

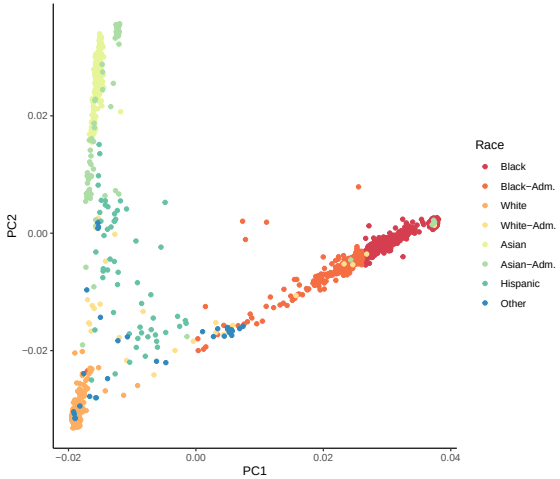
Nephrotic Syndrome association study

- 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

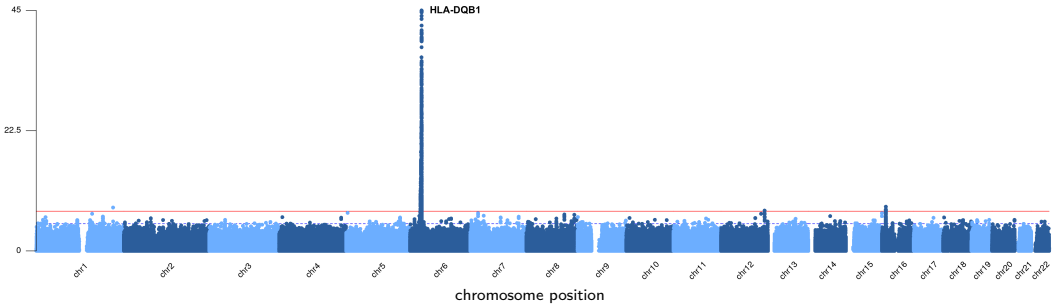
Nephrotic Syndrome association study: Admixture plot



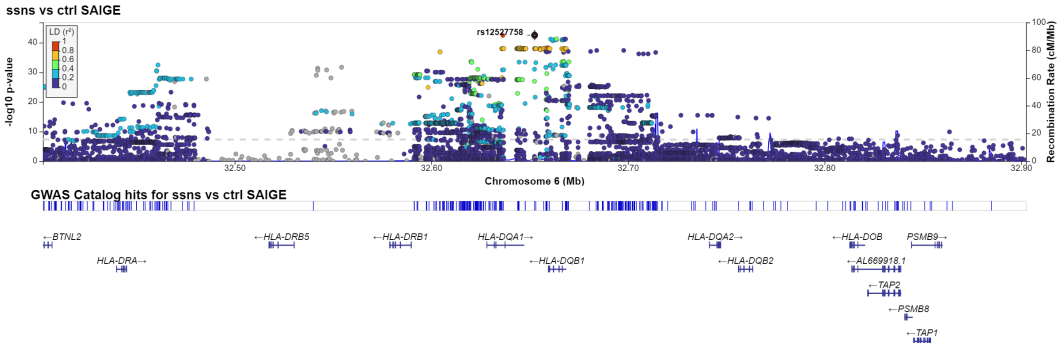
Nephrotic Syndrome association study: PCA plot



Nephrotic Syndrome association study: Manhattan plot

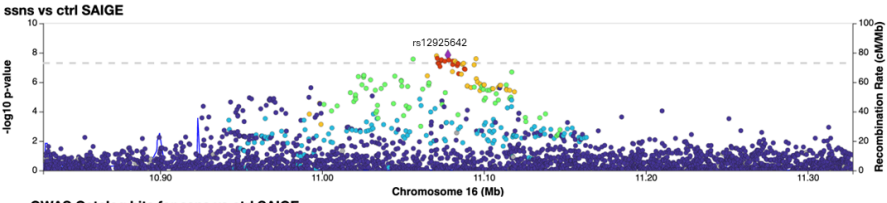


Nephrotic Syndrome: Local plot for HLA region



Many HLA genes in linkage disequilibrium (LD)

Nephrotic Syndrome: Local plot for Chr 16 region

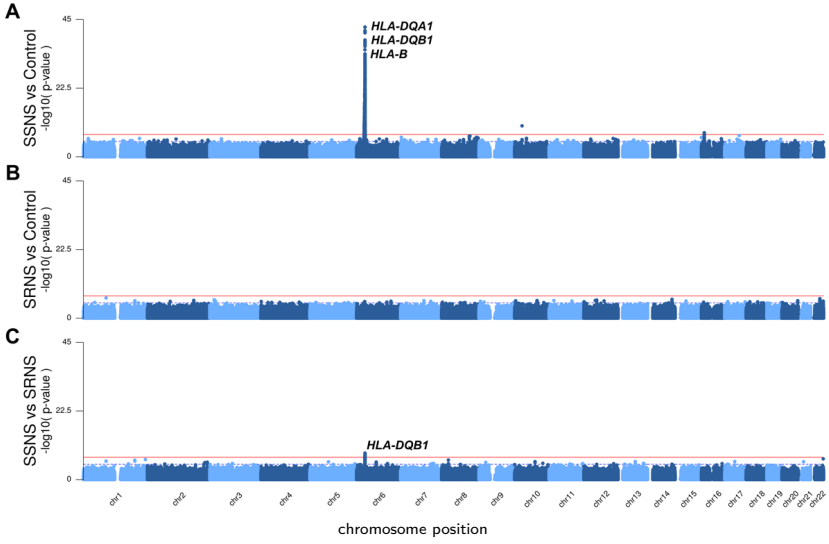


master transcription factor of HLA class II genes

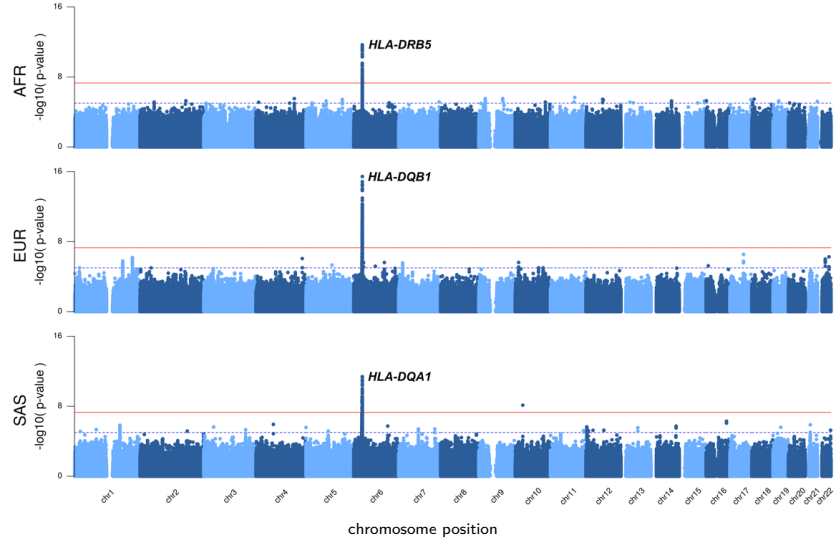
Dexamethasone-induced protein

B cell receptor-dependent HLA-II pathway in human B cells

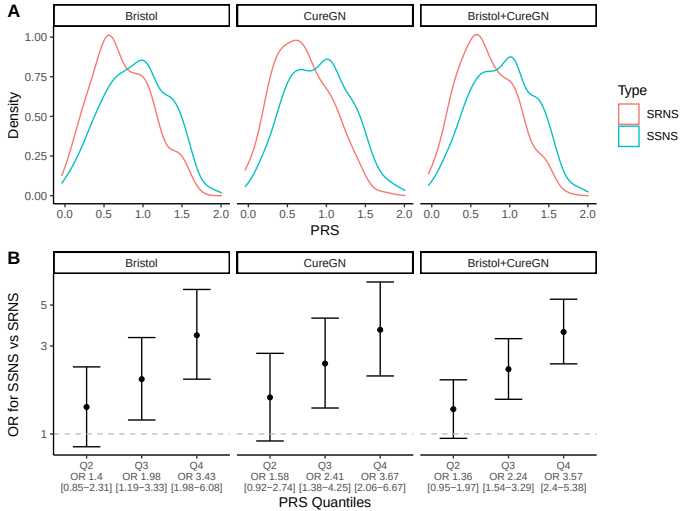
Nephrotic Syndrome: steroid response associated with HLA



Nephrotic Syndrome: HLA association shared across ancestries



Nephrotic Syndrome: PRS help predict steroid response



Acknowledgments

Ochoa Lab
Tiffany Tu
RP Pornmongkolsuk
Zhuoran Hou
Amika Sood
Yiqi Yao

Princeton University
John D. Storey

Duke University
Rasheed Gbadegesin
Kouros Owzar
Beth Hauser
Yi-Ju Li
Andrew Allen
Amy Goldberg

Funding
NIH
Whitehead Scholars

Duke Center for Statistical Genetics and Genomics
Department of Biostatistics & Bioinformatics
Duke University School of Medicine

DrAlexOchoa@genomic.social
ochoalab.github.io
alejandro.ochoa@duke.edu