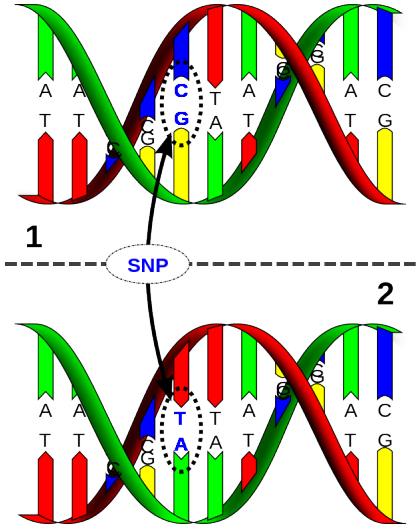
Genetic association models for related samples and population structure

Alejandro Ochoa

Biostatistics and Bioinformatics, StatGen — Duke University

2022-08-26 — BERD Core Seminar

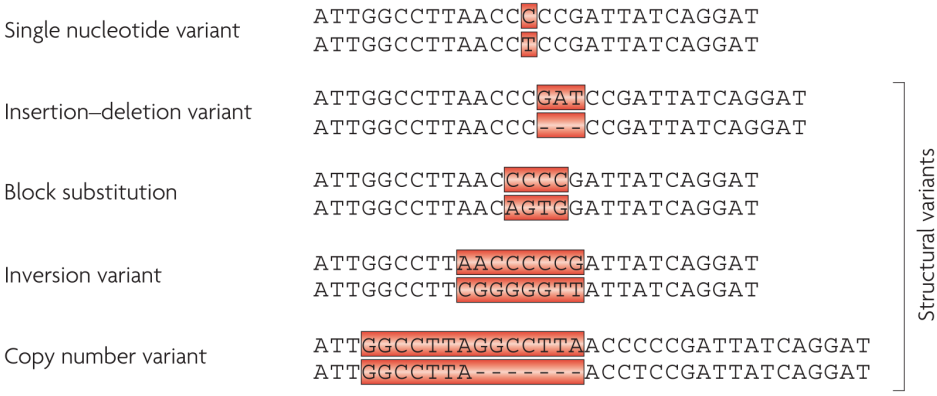
Genetic variation: we're all mutants!



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

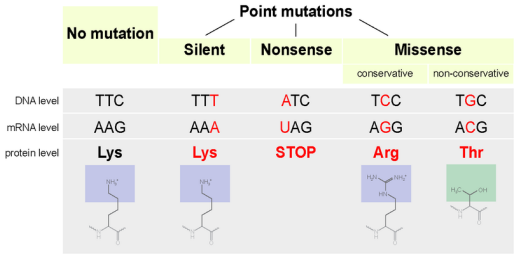


Frazer *et al.* (2009)

- 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!

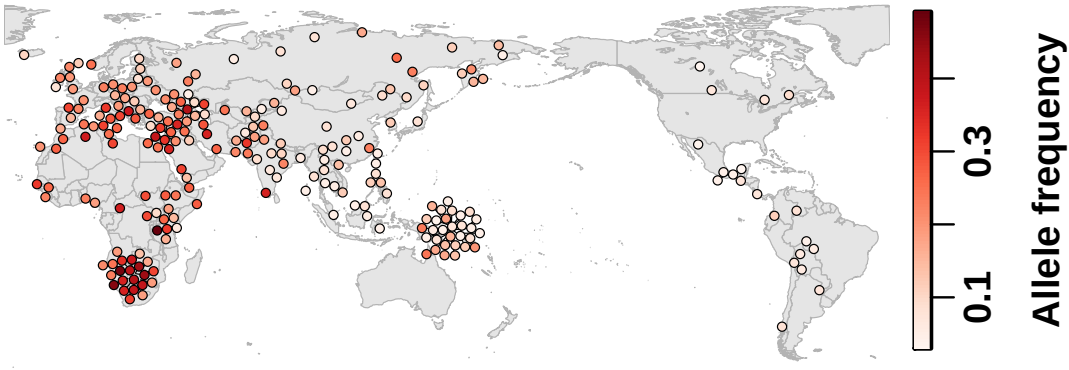
Dynamics of genetic variation

Colors are *alleles*

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

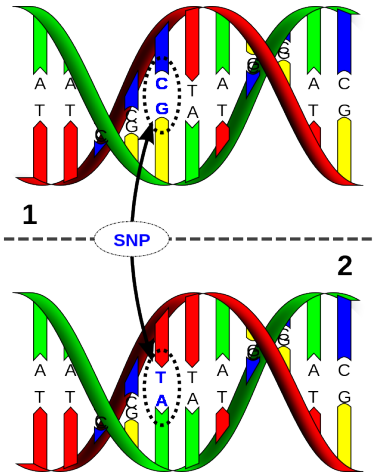


Ochoa and Storey (2019a) doi:10.1101/653279

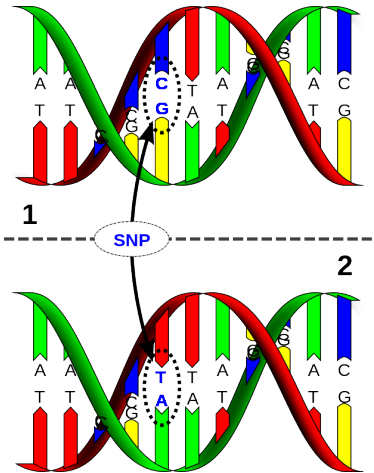
rs17110306; median differentiation given MAF $\geq$ 10%

Why? Migration and isolation, admixture, family structure

Single Nucleotide Polymorphism (SNP) data



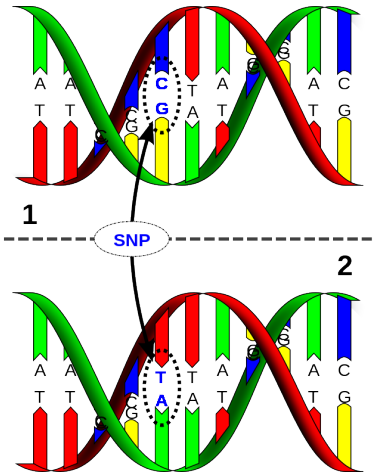
Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

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 ...

$\mathbf{X}$

Hardy-Weinberg Equilibrium (HWE): Binomial draws

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

p_i = frequency of reference allele at locus i.

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:

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

Hardy-Weinberg Equilibrium (HWE): Binomial draws

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

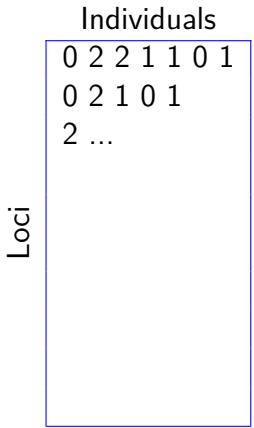
p_i = frequency of reference allele at locus i.

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

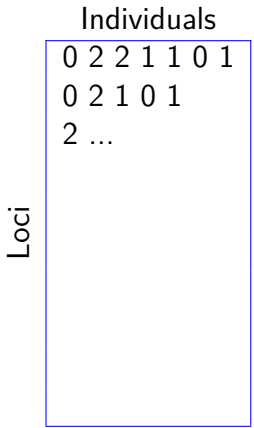


$\mathbf{X}$

High-dimensional binomial data

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

Dependence structure of genotype matrix



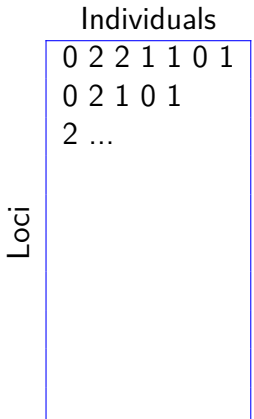
High-dimensional binomial data

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

Relatedness / Population structure

- Dependence between individuals (columns)

Dependence structure of genotype matrix



High-dimensional binomial data

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

Relatedness / Population structure

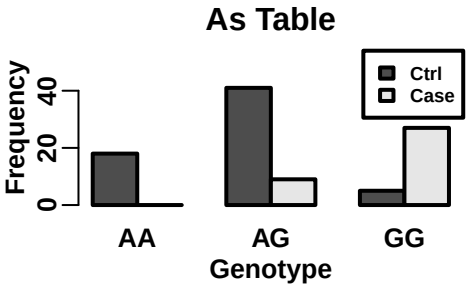
- Dependence between individuals (columns)

Linkage disequilibrium

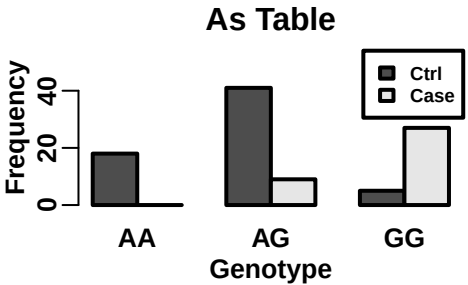
- Dependence between loci (rows)

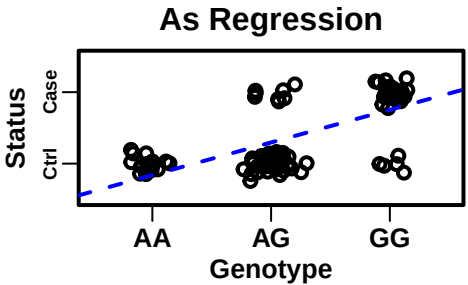
Genetic association study: genotype-phenotype correlation

Genetic association study: genotype-phenotype correlation

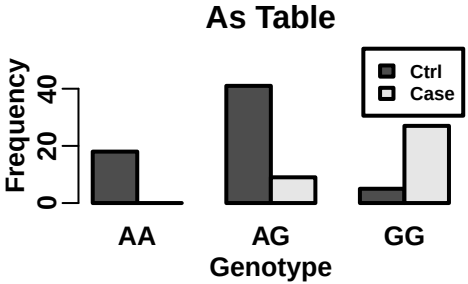


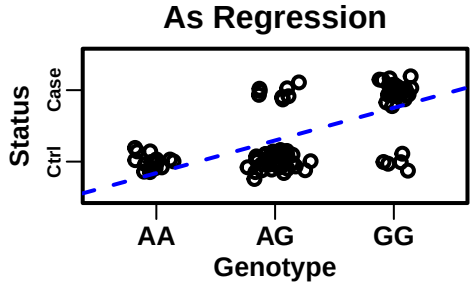
Genetic association study: genotype-phenotype correlation

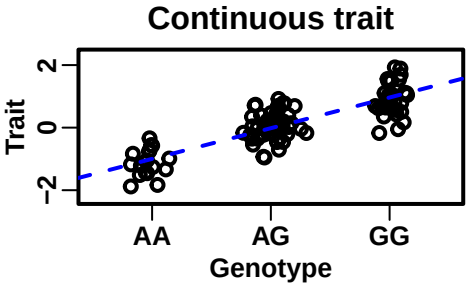




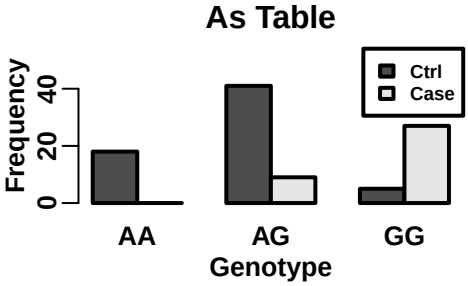
Genetic association study: genotype-phenotype correlation

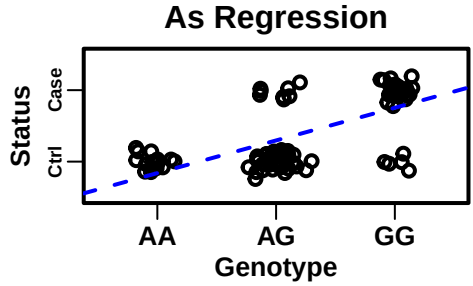


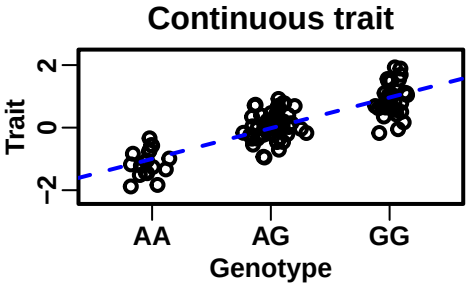


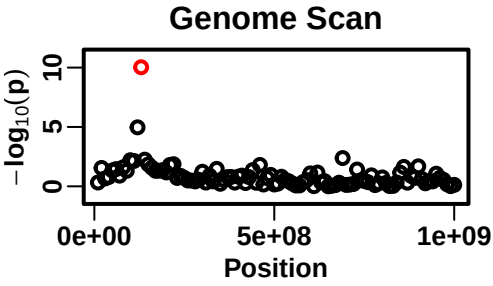


Genetic association study: genotype-phenotype correlation



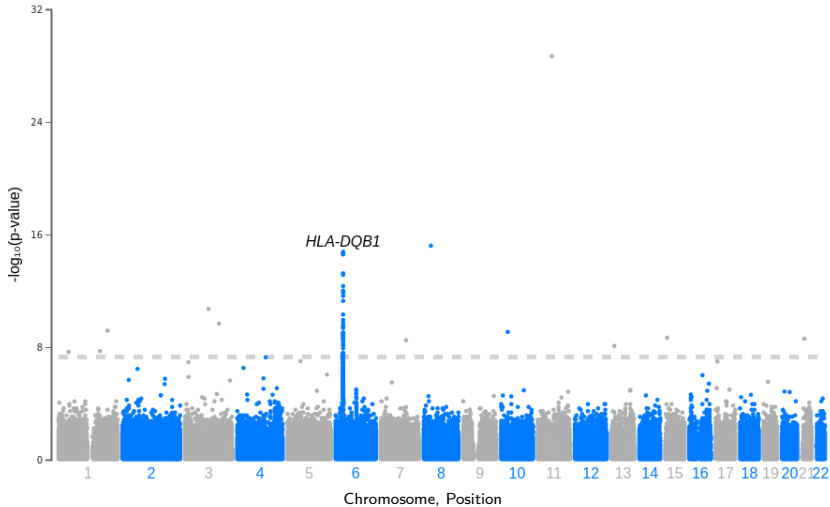




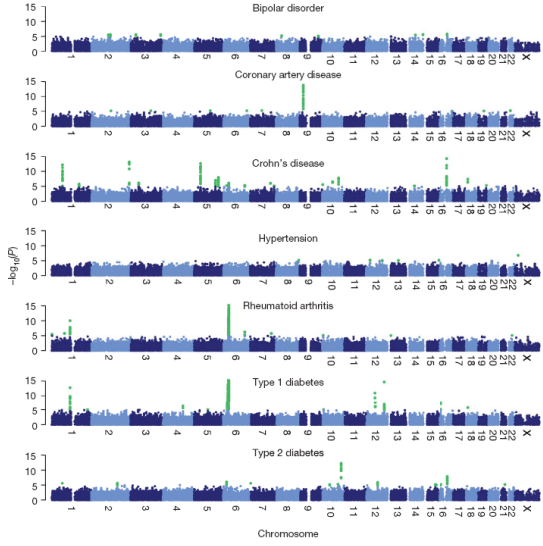


Nephrotic Syndrome association study

Severe pediatric kidney disease. 1000 cases/1000 controls; multiethnic



“Manhattan” plots for other diseases



Wellcome Trust Case Control Consortium (2007)

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.
- But *significant* variants only explain 3% of this heritability.

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.
- But *significant* variants only explain 3% of this heritability.
  - Do we need bigger studies? Some as large as 1M people don't find much!

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.
- But *significant* variants only explain 3% of this heritability.
  - Do we need bigger studies? Some as large as 1M people don't find much!
  - Are most causal variants rare? (causes low statistical power)

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.
- But *significant* variants only explain 3% of this heritability.
  - Do we need bigger studies? Some as large as 1M people don't find much!
  - Are most causal variants rare? (causes low statistical power)
  - Is significance too stringent of a criterion?

This problem is hard!

After the human genome (~2000), researchers thought that was the hard part. Nope!

The *missing heritability* problem:

- Height is highly heritable
  - $h^2 \approx 80\%$: variance explained by genetics, according to *twin/sib studies*.
- But *significant* variants only explain 3% of this heritability.
  - Do we need bigger studies? Some as large as 1M people don't find much!
  - Are most causal variants rare? (causes low statistical power)
  - Is significance too stringent of a criterion?
  - Could it be epigenetic? Shared environment?

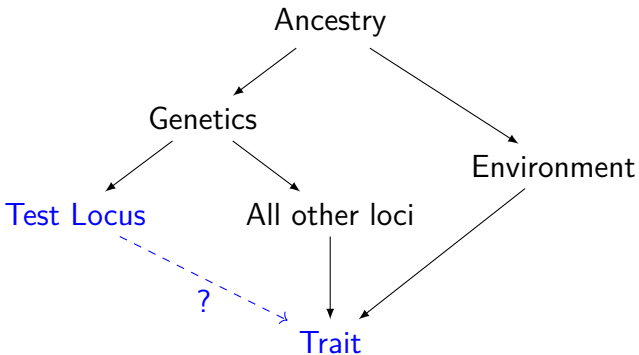
Why is this problem so hard?

Why is this problem so hard?

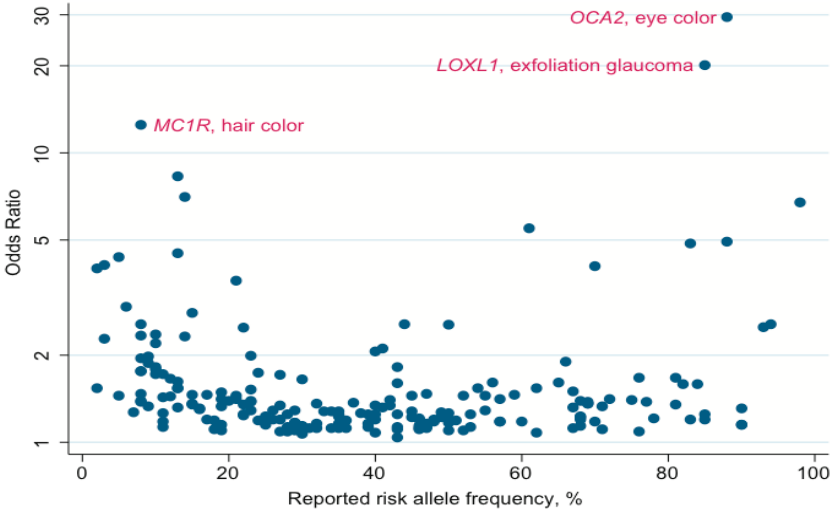
- Millions of tests
- Polygenicity (many causal variants)
- Confounders
- Incorrect assumptions: independence / additivity

Why is this problem so hard?

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



Effects are smaller and rarer than anticipated



Hindorff *et al.* (2009) PNAS 106:9362–9367

Genetic architecture of a trait

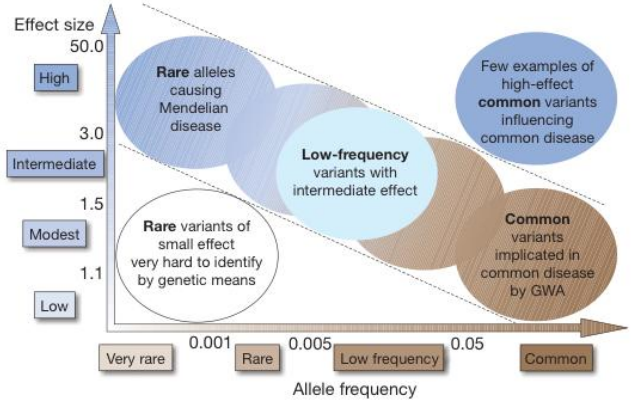


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Manolio *et al.* (2009) Nature 461:747-753

Goal: association, not causation!

- Ideally, we'd actually find the *causal* variants of disease

Goal: association, not causation!

- Ideally, we'd actually find the *causal* variants of disease
- However, causal variants are likely not genotyped

Goal: association, not causation!

- Ideally, we'd actually find the *causal* variants of disease
- However, causal variants are likely not genotyped
- *Linkage Disequilibrium*: variants near the causal locus are correlated to each other and to the disease!

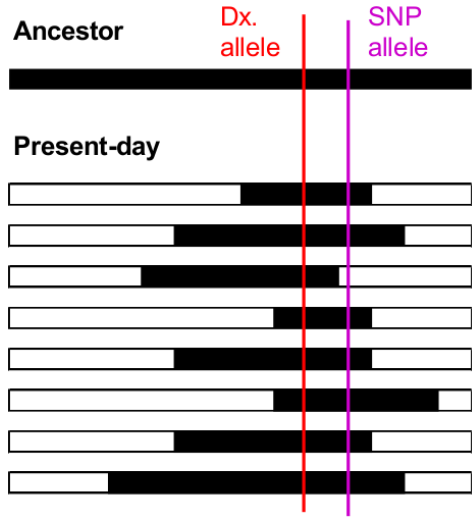


Fig. by Andrew Allen, Duke B&B.

One problem with no causation: prediction outside test pop.

- Association depends on correlation between the tested and causal loci

One problem with no causation: prediction outside test pop.

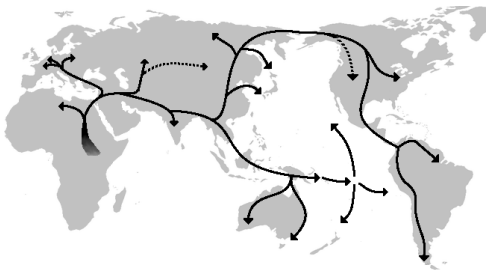
- Association depends on correlation between the tested and causal loci
- But correlation varies in populations! So associations may not be predictive

One problem with no causation: prediction outside test pop.

- Association depends on correlation between the tested and causal loci
- But correlation varies in populations! So associations may not be predictive
- Common scenario:
  - In European-only study, locus i is significantly associated with disease
  - Locus i is not correlated to causal locus in Sub-Saharan Africans
  - So locus i does not predict disease in Sub-Saharan Africans

One problem with no causation: prediction outside test pop.

- Association depends on correlation between the tested and causal loci
- But correlation varies in populations! So associations may not be predictive
- Common scenario:
  - In European-only study, locus i is significantly associated with disease
  - Locus i is not correlated to causal locus in Sub-Saharan Africans
  - So locus i does not predict disease in Sub-Saharan Africans
- Why? Correlations are stronger outside Africa due to population bottleneck

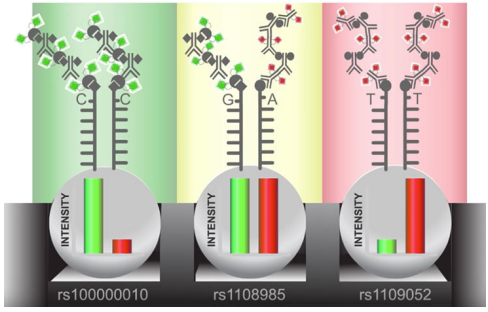


Modern technologies for finding variants

Genotyping arrays vs sequencing

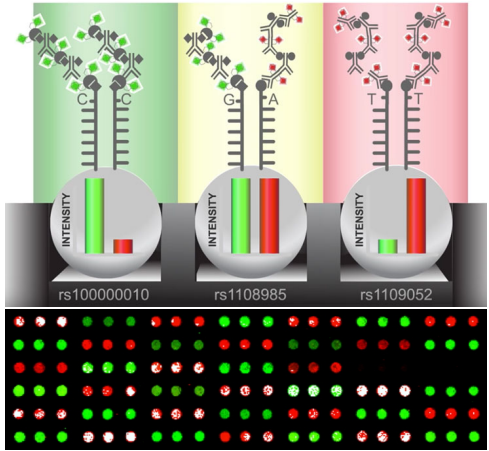
Modern technologies for finding variants

Genotyping arrays vs sequencing



Modern technologies for finding variants

Genotyping arrays vs sequencing



Modern technologies for finding variants

Genotyping arrays

Modern technologies for finding variants

Genotyping arrays

- Oldest and cheapest of the two we discuss here

Modern technologies for finding variants

Genotyping arrays

- Oldest and cheapest of the two we discuss here
- Used by 23andMe, Ancestry, etc.

Modern technologies for finding variants

Genotyping arrays

- Oldest and cheapest of the two we discuss here
- Used by 23andMe, Ancestry, etc.
- Pros:
  - 0.5-1.5 million loci per array
  - Low missingness

Modern technologies for finding variants

Genotyping arrays

- Oldest and cheapest of the two we discuss here
- Used by 23andMe, Ancestry, etc.
- Pros:
  - 0.5-1.5 million loci per array
  - Low missingness
- Cons: tests *known* variants only, a biased set
  - Most often common variants only
  - Preiously: biased for variants common in European ancestry
  - Typically biallelic SNPs (Single Nucleotide Polymorphisms) only
  - Unlikely to contain causal variants
  - Some probes fail $\Rightarrow$ batch effects

Modern technologies for finding variants

Whole genome sequencing

Modern technologies for finding variants

Whole genome sequencing

- Short read sequencing at 2x to 30x depths common

Modern technologies for finding variants

Whole genome sequencing

- Short read sequencing at 2x to 30x depths common
- Variant: whole *exome* sequencing (enriched for protein-coding sequences).

Modern technologies for finding variants

Whole genome sequencing

- Short read sequencing at 2x to 30x depths common
- Variant: whole *exome* sequencing (enriched for protein-coding sequences).
- Pros:
  - More likely to include causal variants
  - Can see short insertions and deletions too (indels)
  - Can *impute* missing data assuming correlations

Modern technologies for finding variants

Whole genome sequencing

- Short read sequencing at 2x to 30x depths common
- Variant: whole *exome* sequencing (enriched for protein-coding sequences).
- Pros:
  - More likely to include causal variants
  - Can see short insertions and deletions too (indels)
  - Can *impute* missing data assuming correlations
- Cons:
  - Still misses repetitive regions, large (structural) variants
  - Need special methods for rare variants
  - More expensive (for now)

Modern technologies for finding variants

	Microarrays	Whole genome seq
Cost/person (2019)	\$50-100	\$700-1000
Loci	0.5-1.5 M (fixed)	up to 80 M ? (random)
Missingness	Low	High
Causal locus tested?	Probably no	Probably yes

Population structure: lack of independence between individuals

Population structure: lack of independence between individuals

In classical association studies, every individual is treated as *independent*.

Population structure: lack of independence between individuals

In classical association studies, every individual is treated as *independent*.

In a case-control study, we test for a bias in allele frequencies (X is a random genotype):

$$\begin{aligned} X|\text{case} &\sim \text{Binomial}(2, p_{\text{case}}), \\ X|\text{control} &\sim \text{Binomial}(2, p_{\text{control}}), \\ \text{reject } H_0 \text{ if:} \quad & p_{\text{case}} \neq p_{\text{control}}. \end{aligned}$$

Population structure: lack of independence between individuals

In classical association studies, every individual is treated as *independent*.

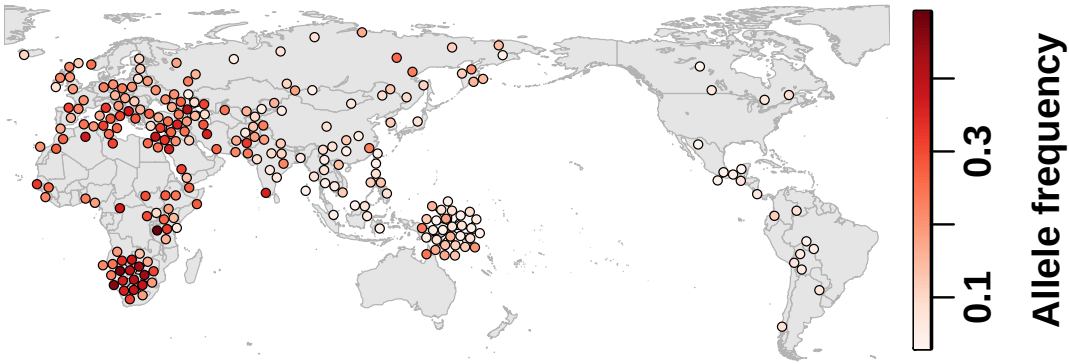
In a case-control study, we test for a bias in allele frequencies (X is a random genotype):

$$\begin{aligned} X|\text{case} &\sim \text{Binomial}(2, p_{\text{case}}), \\ X|\text{control} &\sim \text{Binomial}(2, p_{\text{control}}), \\ \text{reject } H_0 \text{ if:} \quad & p_{\text{case}} \neq p_{\text{control}}. \end{aligned}$$

However:

- Allele frequencies often vary between human subpopulations
- Disease prevalence may also vary between subpopulations (if causal loci also vary in frequency across the world!)

Median human locus by differentiation

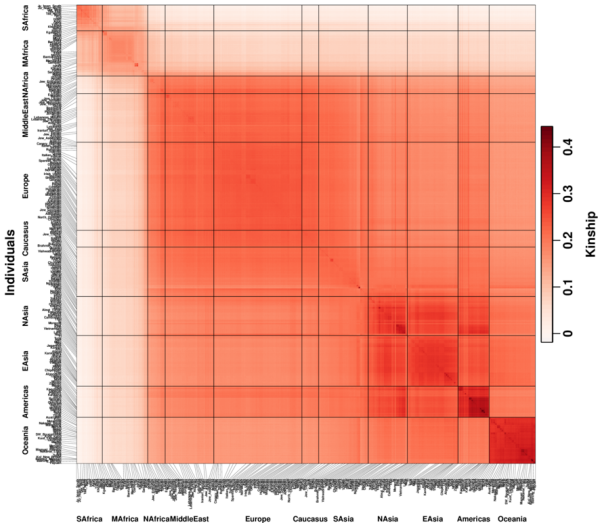


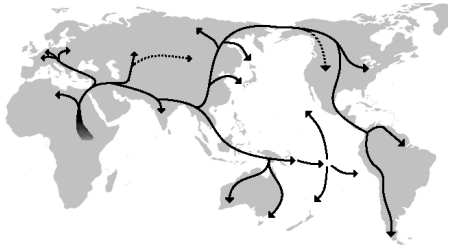
Ochoa and Storey (2019a) doi:10.1101/653279

rs17110306; median differentiation among loci with minor allele frequency $\geq 10\%$

Classical association tests assume allele frequency is the same across the world!

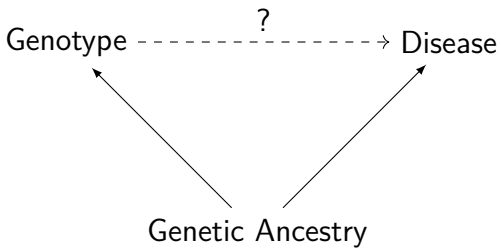
Kinship (covariance) matrix of world-wide human population



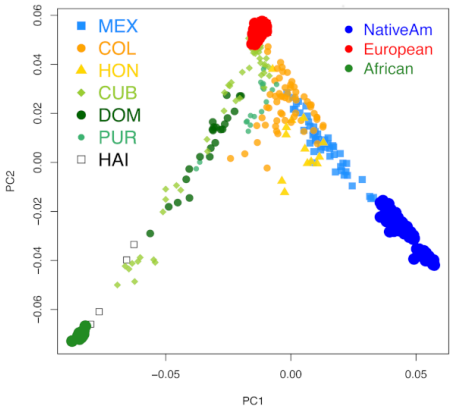


Ochoa and Storey (2019) doi:10.1101/653279

Ancestry as a statistical confounder



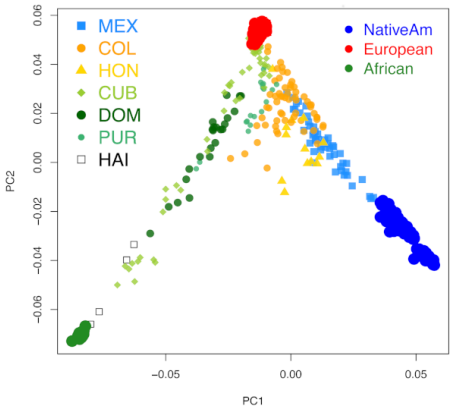
PCA: Principal Component Analysis



Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

PCA: Principal Component Analysis

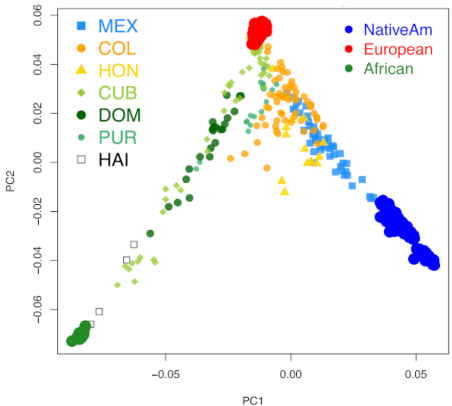


Moreno-Estrada *et al.* (2013)

Use top eigenvectors of covariance matrix in any regression approach!

PCs map to ancestry.

PCA: Principal Component Analysis



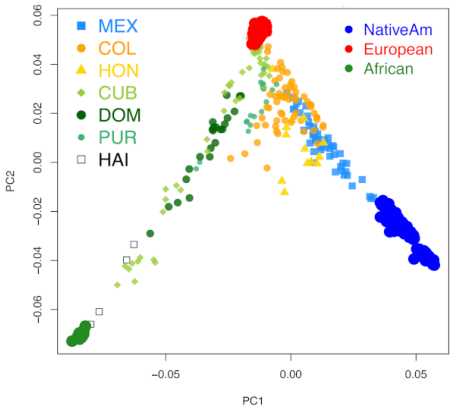
Moreno-Estrada *et al.* (2013)

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.

Genetic association for structured pops: PCA and LMM

Genetic association for structured pops: PCA and LMM

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

Genetic association for structured pops: PCA and LMM

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

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

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

Genetic association for structured pops: PCA and LMM

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

Genetic association for structured pops: PCA and LMM

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

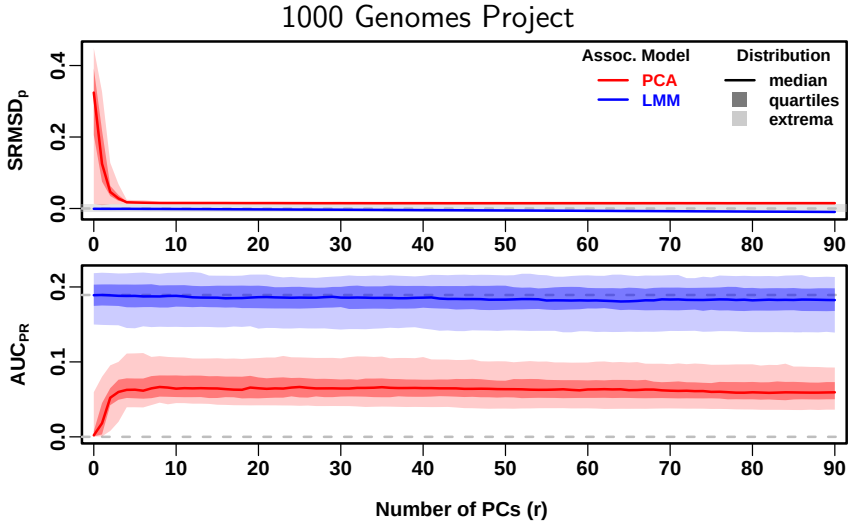
$$\text{PCA}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{U}_d\gamma_d + \epsilon,$$
$$\text{LMM}: \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta + \mathbf{s} + \epsilon.$$

$\mathbf{U}_d$ 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

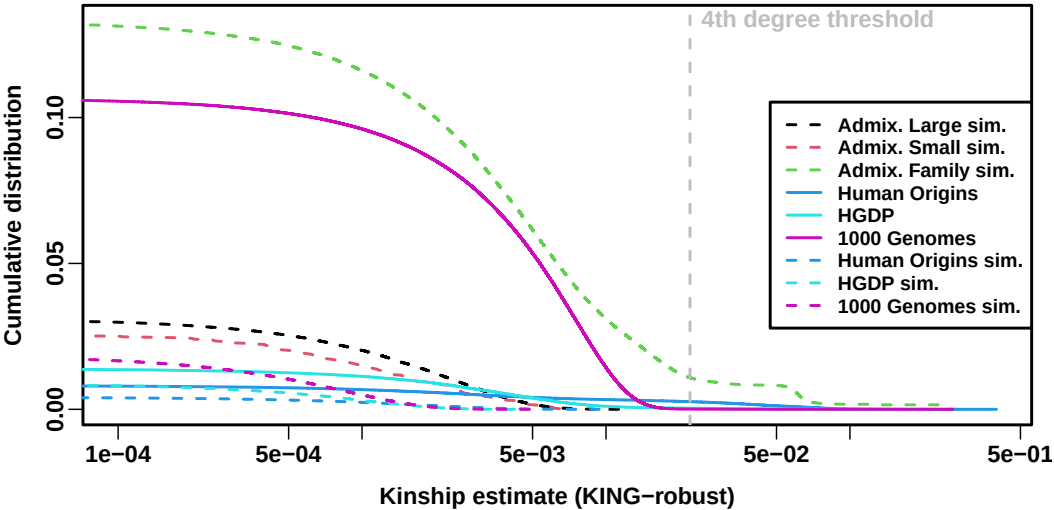
Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

PCA < LMM in association for real datasets



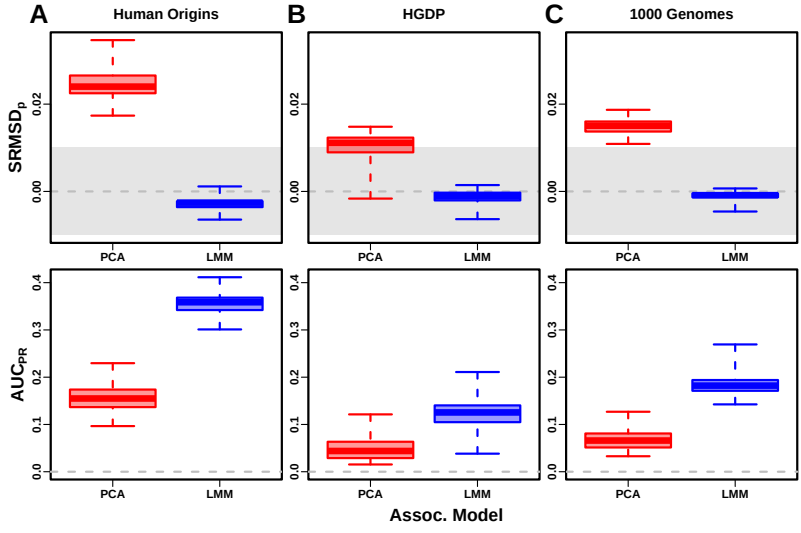
Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

Numerous distant relatives in real datasets



Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

Numerous distant relatives in real datasets explain PCA < LMM



Yao and Ochoa (2022) doi:10.1101/2022.03.25.485885

What happens after we find significant loci?

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"
- Validate experimentally (animal model, tissue culture)

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"
- Validate experimentally (animal model, tissue culture)

Association variants are hard to interpret without experiments:

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"
- Validate experimentally (animal model, tissue culture)

Association variants are hard to interpret without experiments:

- 3% of human genome is protein-coding (most interpretable)

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"
- Validate experimentally (animal model, tissue culture)

Association variants are hard to interpret without experiments:

- 3% of human genome is protein-coding (most interpretable)
- Most non-coding sequences are of unknown function
  - Except: promoters, enhancers, splice sites, etc

What happens after we find significant loci?

Recall we probably do not have causal locus (unless using deep sequencing).

- Verify association in a *validation* dataset (disjoint from initial study)
- Fine mapping: sequence region and retest
  - Beware "winner's curse"
- Validate experimentally (animal model, tissue culture)

Association variants are hard to interpret without experiments:

- 3% of human genome is protein-coding (most interpretable)
- Most non-coding sequences are of unknown function
  - Except: promoters, enhancers, splice sites, etc
- Link intergenic variant to closest gene often incorrect!

What happens after we find significant loci?

- ... and then, variant/gene *might* suggest a treatment the disease

What happens after we find significant loci?

- ... and then, variant/gene *might* suggest a treatment the disease
- Test treatment in vitro 🧪

What happens after we find significant loci?

- ... and then, variant/gene *might* suggest a treatment the disease
- Test treatment in vitro 🧪
- Test on an animal model 🐁

What happens after we find significant loci?

- ... and then, variant/gene *might* suggest a treatment the disease
- Test treatment in vitro 🧪
- Test on an animal model 🐁
- Test on humans 🧑

What happens after we find significant loci?

- ... and then, variant/gene *might* suggest a treatment the disease
- Test treatment in vitro 🧪
- Test on an animal model 🐁
- Test on humans 🧑
- Make money 💵