

# Genetics and Evolution

---

CHARLESTON CHIANG, PH.D.

BISC 577

12.3.2019

# Disclaimers and Attributions

---

It is obviously impossible to cram everything to do with “Genetics and Evolution” into a lecture. I will not do justice to any topic I cover. Thus I will lightly touch upon a number of main themes, hoping to interest you to seek more information.

The lecture (and myself) had been heavily influenced by the work and teaching of Alkes Price, Graham Coop, and John Novembre.

# What is Population Genetics?

---

**Population genetics** is the study of genetic variation, both within and between (human) populations.

It is the study of population genetic mechanisms (**mutations, assortative mating, migration, drift, recombination, and selection, etc.**) and how these forces shape the pattern of genetic variation.

*Nothing in biology makes sense except in light of evolution.* –T. Dobzhansky (1973)

*Nothing in evolution makes sense except in light of population genetics.* –M. Lynch (2005)

# Are different human populations actually genetically different?

---



<https://www.universiteitleiden.nl/en/research-dossiers/language-diversity>

Slightly.

5-7% of worldwide human genetic variation is due to genetic differences between human populations.

The remaining 93-95% of human genetic variation is due to genetic variation **within** human populations  
(Rosenberg et al. 2002 Science)

# Does “race” exist?

---

As Europeans explored and colonised the world [over the last few centuries], thinkers, philosophers and scientists from those countries attempted to apply taxonomic structures to the people that they encountered, and though these attempts were many and varied, they typically reflected sharp geographic boundaries, and obvious physical characteristics, such as pigmentation and basic morphology – that is to say, what people look like.

--Birney, Raff, Rutherford, and Scally. 2019

- <http://ewanbirney.com/2019/10/race-genetics-and-pseudoscience-an-explainer.html>

# Does “race” exist? Not by genetics

---

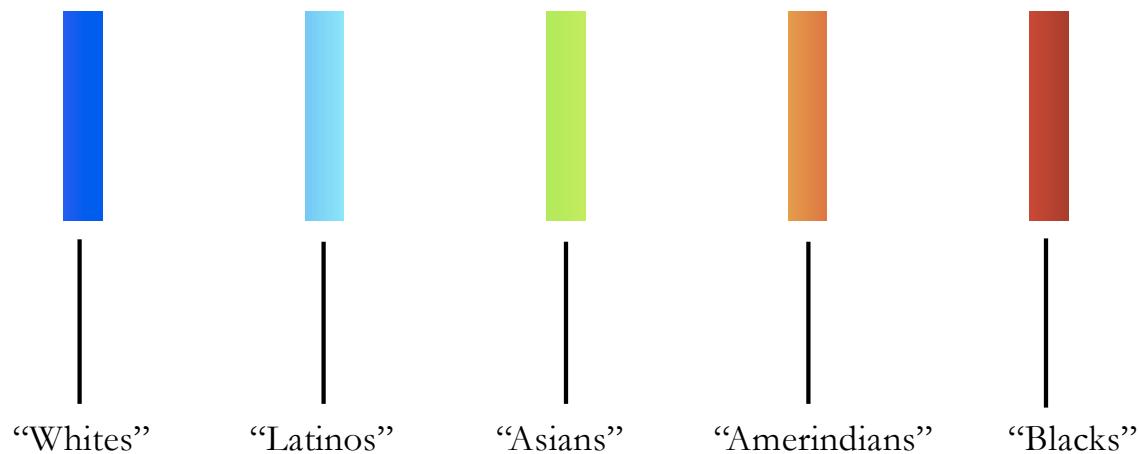
Racial classifications are inadequate descriptors of the distribution of human genetic variation. (Tishkoff & Kidd, 2004 Nat. Genet.)

World-wide patterns of human genetic variation are best described using continuous clines instead of discrete clusters. (Serre & Paabo, 2004 Genome Res.)

If an alien, arriving on Earth with no knowledge of our social history, wished to categorise human ancestry purely on the basis of genetic data, they would find that any consistent scheme must include many distinct groups within Africa that are just as different from each other as Africans are to non-Africans. (Birney et al.)

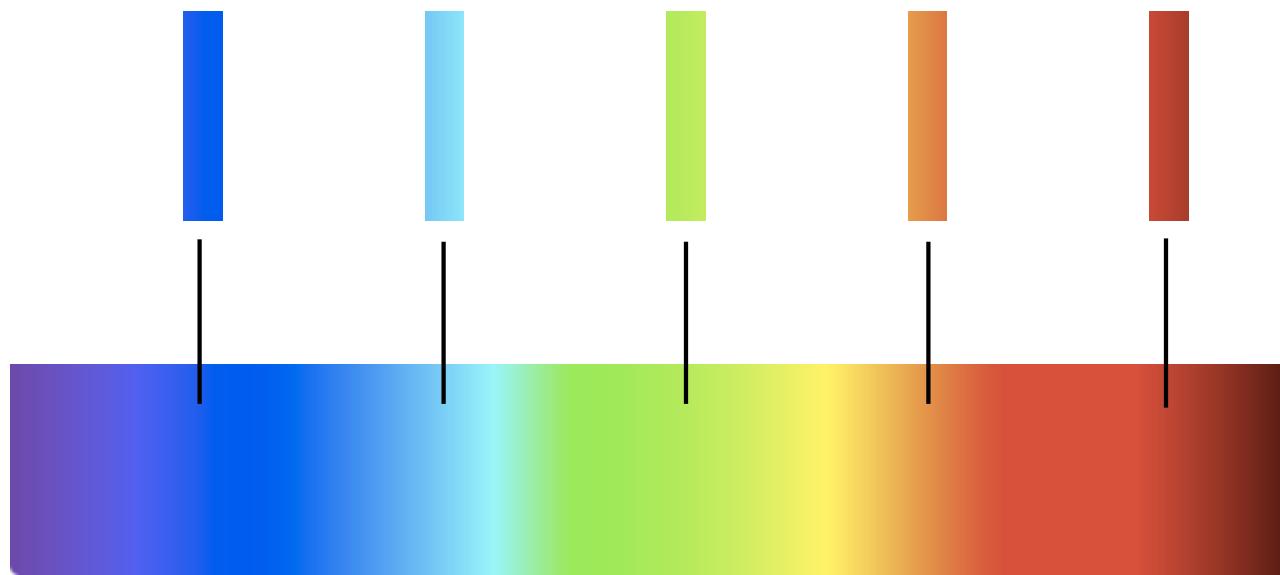
# Does “race” exist? Not by genetics

---



# Does “race” exist? Not by genetics

---

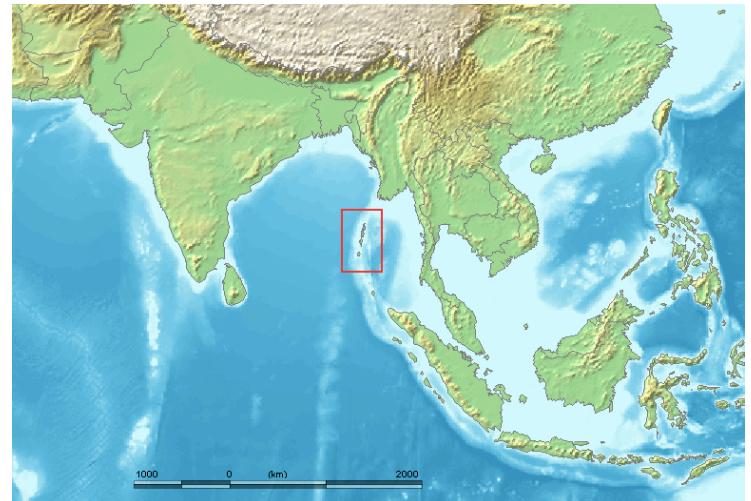


# What “race” are these individuals?

---



Onge population, one of the Andamanese indigenous people of the Andaman Islands



Of course,  
sometimes  
geneticists  
aren't  
helping...

## Multiethnic Genome-Wide Association Study of Diabetic Retinopathy Using Liability Threshold Modeling of Duration of Diabetes and Glycemic Control

*Diabetes* 2019;68:441–456 | <https://doi.org/10.2337/db18-0567>

To identify genetic variants associated with diabetic retinopathy (DR), we performed a large multiethnic genome-wide association study. Discovery included eight European cohorts ( $n = 3,246$ ) and seven African American cohorts ( $n = 2,611$ ). We meta-analyzed across cohorts using inverse-variance weighting, with and without liability threshold modeling of glycemic control and duration of diabetes. Variants with a  $P$  value  $<1 \times 10^{-5}$  were investigated in replication cohorts that included 18,545 European, 16,453 Asian, and 2,710 Hispanic subjects. After correction for multiple testing, the C allele of rs142293996 in an intron of nuclear VCP-like (*NVL*) was associated with DR in European discovery cohorts ( $P = 2.1 \times 10^{-9}$ ), but did not reach genome-wide significance after meta-analysis with replication cohorts. We applied

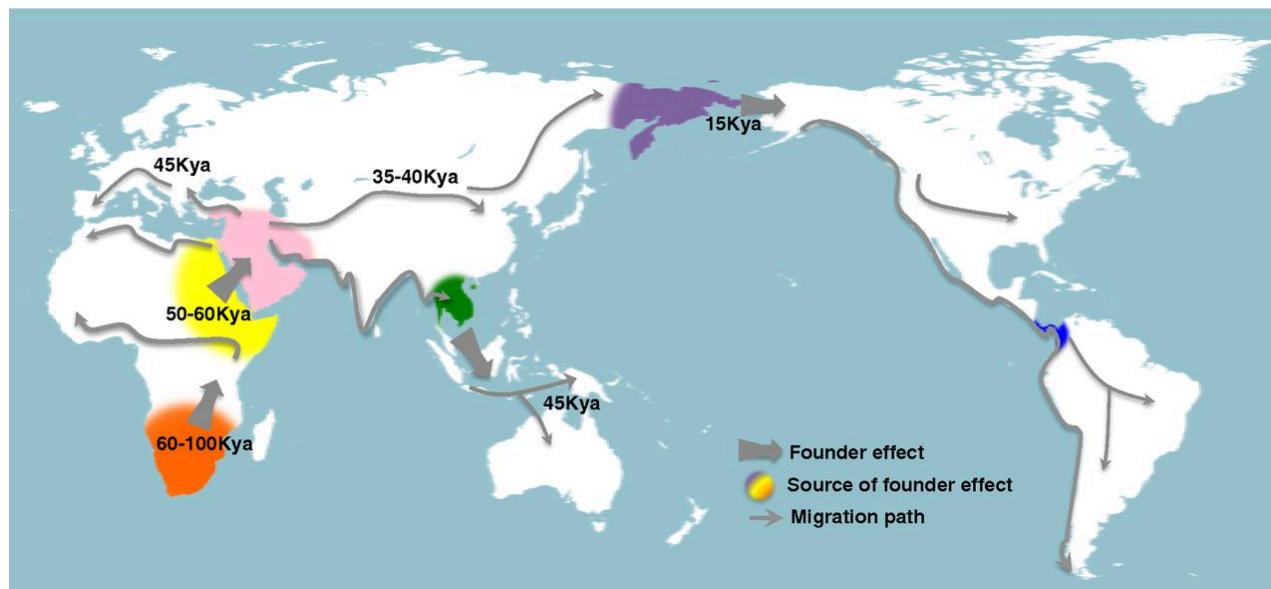
the Disease Association Protein-Protein Link Evaluator (DAPPLE) to our discovery results to test for evidence of risk being spread across underlying molecular pathways. One protein-protein interaction network built from genes in regions associated with proliferative DR was found to have significant connectivity ( $P = 0.0009$ ) and corroborated with gene set enrichment analyses. These findings suggest that genetic variation in *NVL*, as well as variation within a protein-protein interaction network that includes genes implicated in inflammation, may influence risk for DR.

Diabetic retinopathy (DR) is a leading cause of blindness (1). Established risk factors include longer duration of

# Why study differences between human populations?

---

Learn about human migration patterns and ancient history



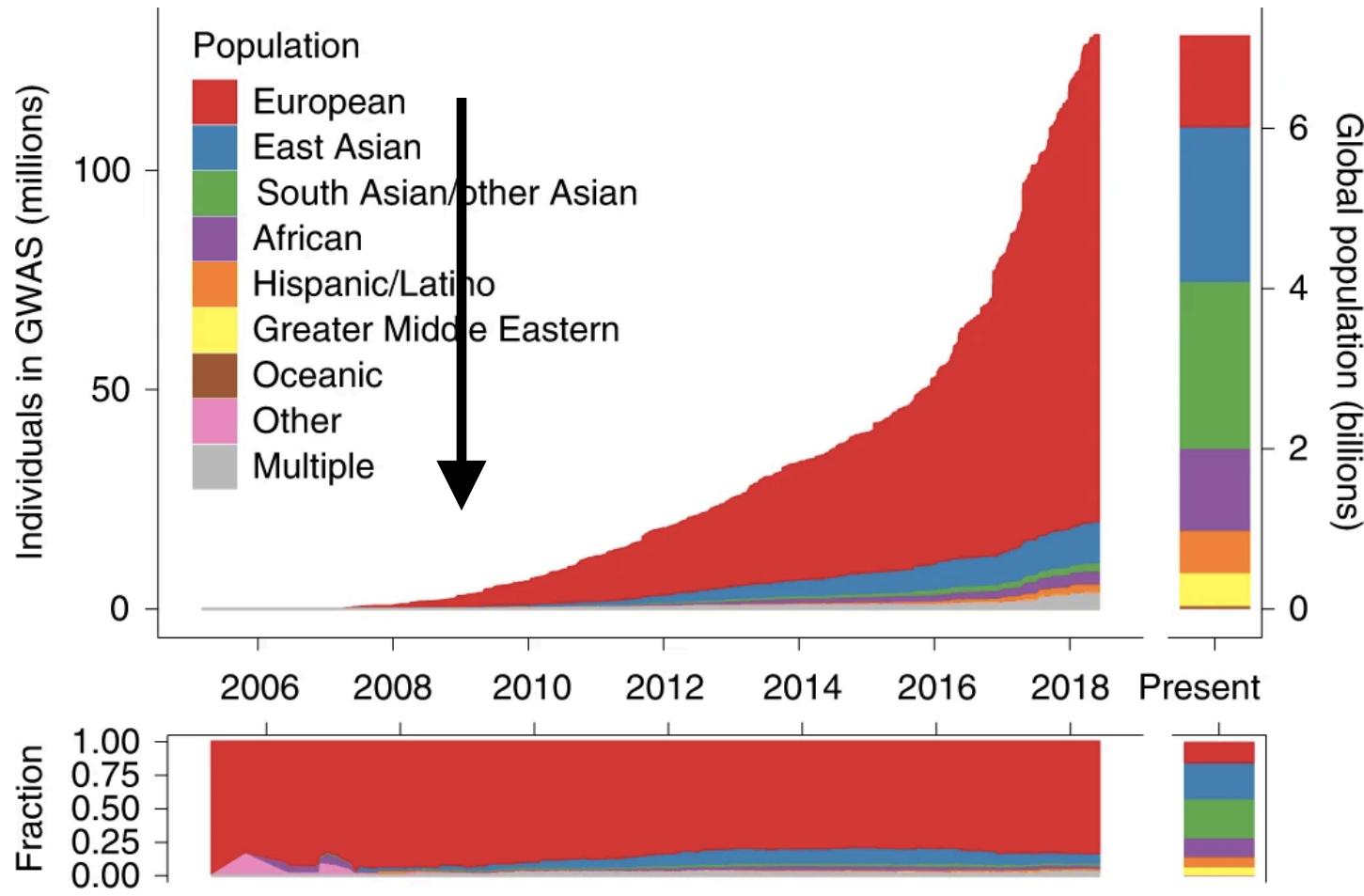
Henn et al. PNAS 2012

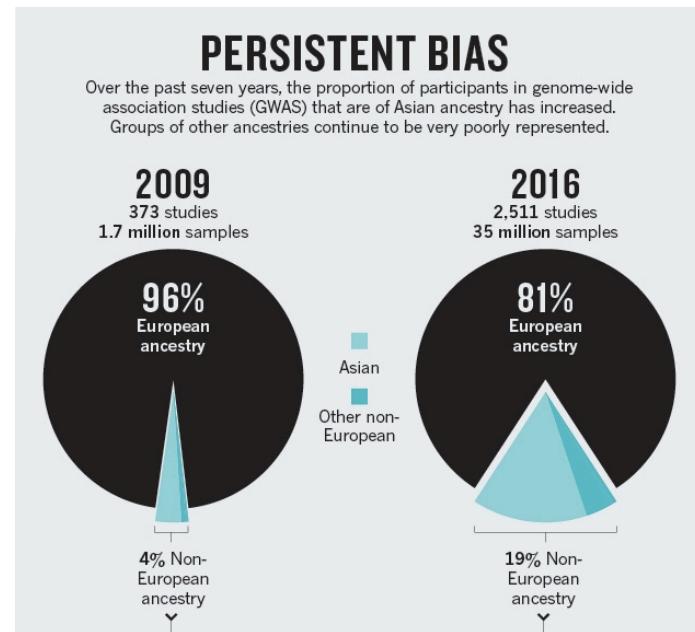
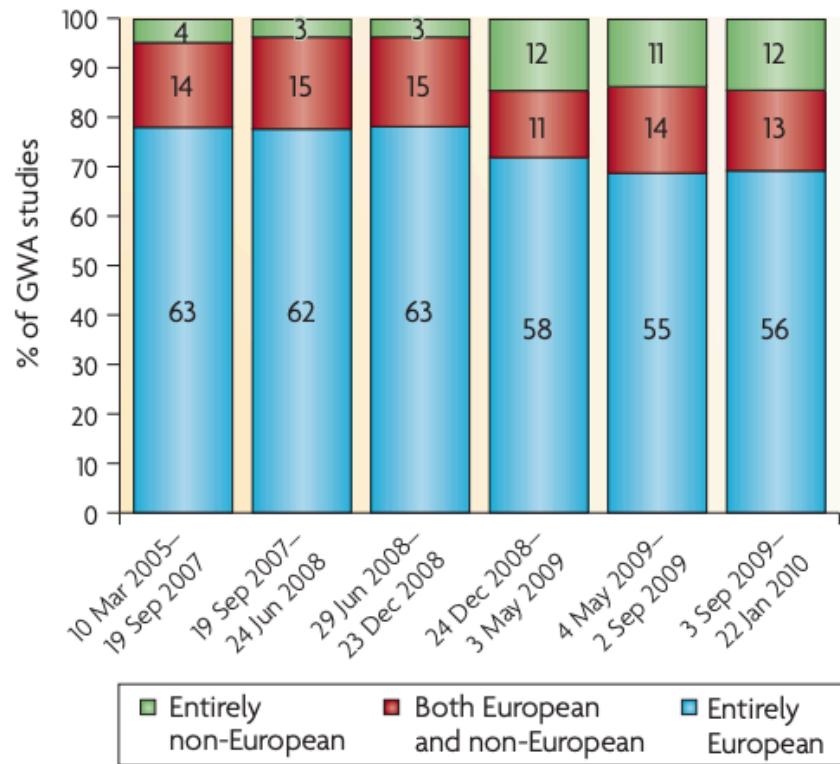
# Why study differences between human populations?

---

Learn about human migration patterns and ancient history

Improve our power to identify and localize disease genes





Rosenberg et al. Nat. Rev. Genet. 2010

Popejoy & Fullerton, Nature 2016

LETTERS

nature  
genetics

---

## A thrifty variant in *CREBRF* strongly influences body mass index in Samoans

Ryan L Minster<sup>1,13</sup>, Nicola L Hawley<sup>2,13</sup>, Chi-Ting Su<sup>1,12,13</sup>, Guangyun Sun<sup>3,13</sup>, Erin E Kershaw<sup>4</sup>, Hong Cheng<sup>3</sup>, Olive D Buhule<sup>5,12</sup>, Jerome Lin<sup>1</sup>, Muagututi'a Sefuiva Reupena<sup>6</sup>, Satupa'itea Viali<sup>7</sup>, John Tuitele<sup>8</sup>, Take Naseri<sup>9</sup>, Zsolt Urban<sup>1,14</sup>, Ranjan Deka<sup>3,14</sup>, Daniel E Weeks<sup>1,5,14</sup> & Stephen T McGarvey<sup>10,11,14</sup>

## LETTER

---

---

doi:10.1038/nature12828

## Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico

The SIGMA Type 2 Diabetes Consortium\*

# Why study differences between human populations?

---

Learn about human migration patterns and ancient history

Improve our power to identify and localize disease genes

- Differential patterns in linkage disequilibrium helps with fine-mapping.
- Avoid false positives due to population stratification.
- Signals of natural selection at genes related to disease.

# Why study differences between human populations?

---

Learn about human migration patterns and ancient history

Improve our power to identify and localize disease genes

- Differential patterns in linkage disequilibrium helps with fine-mapping.
- Avoid false positives due to population stratification.
- Signals of natural selection at genes related to disease.

Improve health disparity, promote personalized medicine

# Lecture Outline

---

1. ~~Introduction and Motivation~~
2. Pattern of Genetic Variations
3. Population Genetic Forces that Impacts Genetic Variations
  - Demography
  - Natural Selection

# What is genetic variation?

---

Genetic variation is the **difference in DNA** sequences between individuals. Variation occurs in **germ cells** (i.e. sperm and egg), and also in somatic cells.

Mutations and recombination are major sources of variation

There are many types of genetic variation:

- **Single Nucleotide Polymorphisms (SNP)** or Single Nucleotide Variants (SNV) ...
- Structural Variants (SV): Insertions, Deletions, Inversions, Copy Number Variants (CNV), Duplications ...
- Short Tandem Repeats (STR)
- Large scale chromosomal rearrangements ...

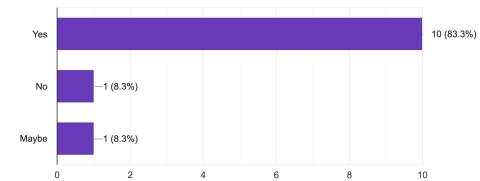
# What is a Single Nucleotide Polymorphism (SNP)?

---

It is a letter of the genetic code at a particular genomic position that differs in different individuals (e.g. chromosome 1, base pair 50,055,936, G/T).

1. Do you know what is a SNP ?

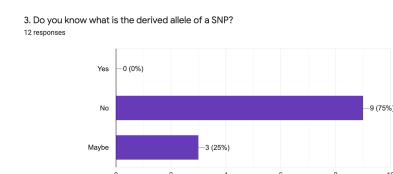
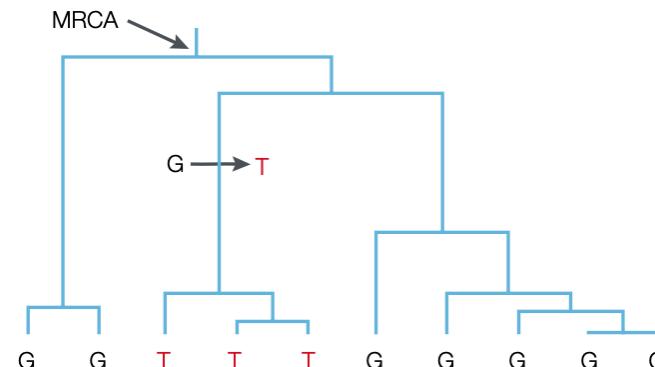
12 responses



# What is a Single Nucleotide Polymorphism (SNP)?

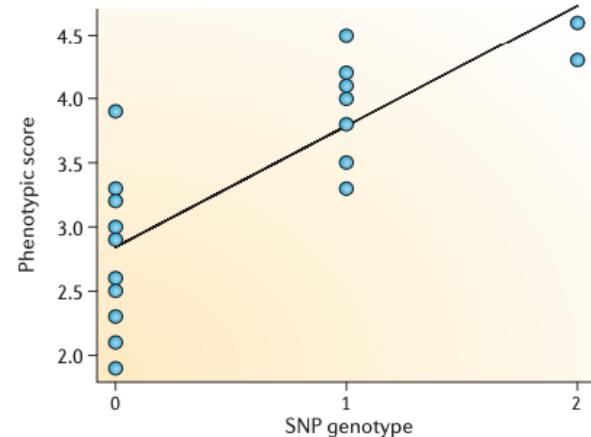
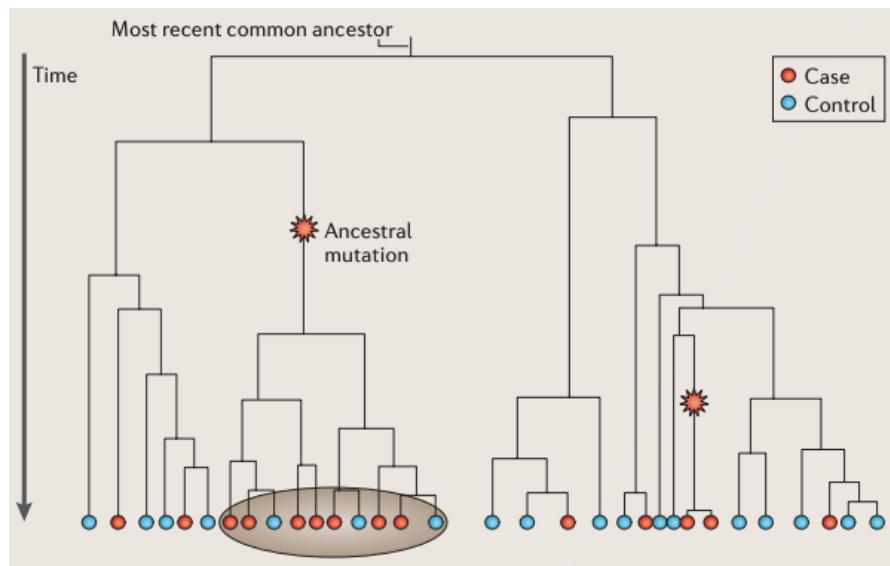
It is a letter of the genetic code at a particular genomic position that differs in different individuals (e.g. chromosome 1, base pair 50,055,936, G/T).

Each SNP is (typically assumed) to correspond to a single mutation event in history, e.g. G mutated to T in a single ancestor. Then G = **ancestral** allele, T = **derived** allele.



Rosenberg & Nordborg, Nat. Rev. Genet. 2002

# Evolutionary Rationale for GWAS

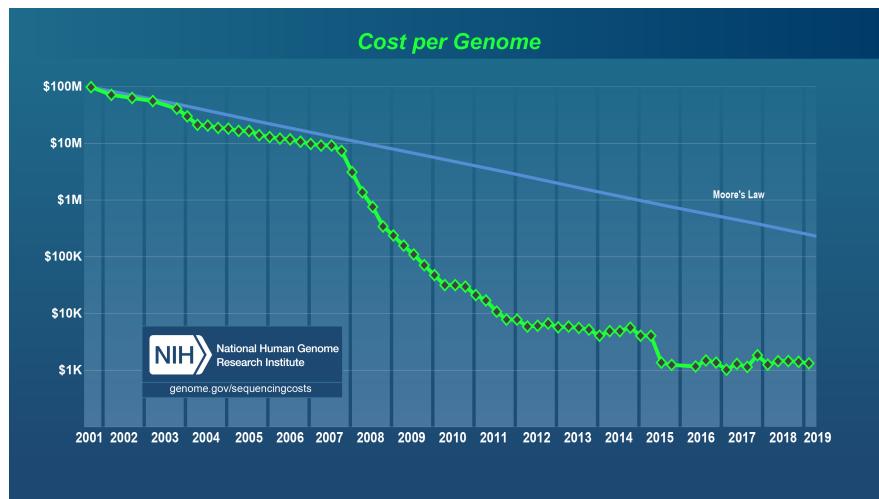


Balding, Nat. Rev. Genet. 2006

# How do we ascertain genetic variation?

---

Sequencing (whole genome or whole exome), but until this day it is still not feasible at large scale.



Moore's Law is the empirical relationship that number of transistors in a dense integrated circuit doubles about every two years. It's a projection of increase in computer power and decrease in relative cost.

# How do we ascertain genetic variation?

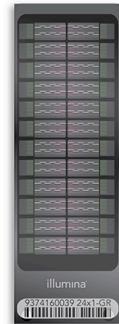
---

Sequencing (whole genome or whole exome), but until this day it is still not feasible at large scale.

Historically (and still now), we relied on genotyping microarrays to genotype specific locations in the human genome known to harbor SNP variation.



Affymetrix



Illumina

# How do we ascertain genetic variation?

---

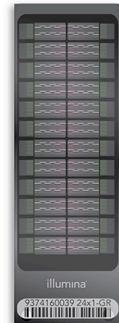
Sequencing (whole genome or whole exome), but until this day it is still not feasible at large scale.

Historically (and still now), we relied on genotyping microarrays to genotype specific locations in the human genome known to harbor SNP variation.

Thus, we relied on large database of SNP variation.



Affymetrix



Illumina



# The International HapMap Consortium

Phase 1: Launched in 2002, completed in 2005, targeted > 1M SNPs in the genome for genotyping in 4 populations (269 samples).

Some SNP discovery efforts (shot-gun sequencing in limited samples), combined with dbSNP and whatever information available at the time. Then designed SNP assays in 5kb bins at a time across the genome. Limited use of arrays (40K and 120K)

Population Label	Population Name	Sample Size
CEU	Utah residents with Northern and Western European ancestry from CEPH collection	90
CHB	Han Chinese in Beijing, China	45
JPT	Japanese in Tokyo, Japan	44
YRI	Yoruba in Ibadan, Nigeria	90



# The International HapMap Consortium

---

Phase 2: Completed in 2007, driven by Perlegen custom array (4.3M assays), but also Affymetrix 500K and Illumina 100K and 300K arrays.

3.1M SNPs passed QC and polymorphic in at least 1 population. 2.8M SNPs with minor allele frequency > 0.05 in at least 1 population (“common”).

Population Label	Population Name	Sample Size
CEU	Utah residents with Northern and Western European ancestry from CEPH collection	90
CHB	Han Chinese in Beijing, China	45
JPT	Japanese in Tokyo, Japan	44
YRI	Yoruba in Ibadan, Nigeria	90



# The International HapMap Consortium

Phase 3: Completed in 2010, only used Affymetrix 6.0 (900K SNPs) and Illumina Infinium 1M (1M SNPs). 99.8% call rate, 99.5% concordance.

Together 1.6M SNPs, but also tap into rarer variants, CNVs, and much more diverse populations (N = 1184).

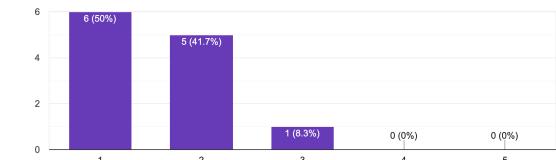
Pop Label	Population Name	Sample Size
ASW	African ancestry Southwest USA	83
CEU	Utah resident, N & W European ancestry	165
CHB	Han Chinese in Beijing, China	84
CHD	Chinese in Metropolitan Denver, CO	85
GIH	Gujarati Indians in Houston, TX	88
JPT	Japanese in Tokyo, Japan	86
LWK	Luhya in Webuye, Kenya	90
MXL	Mexican ancestry in Los Angeles, CA	77
MKK	Maasai in Kinyawa, Kenya	171
TSI	Toscani in Italia	88
YRI	Yoruba in Ibadan, Nigeria	167

# HapMap Project Summary

Project	HapMap		
Phase	1	2	3
Year Complete	2005	2007	2010
Sample Size	269	270	1184
Populations	4	4	11
# variants	1 M	3.1 M	1.6 M + CNVs

6. How familiar are you with the HapMap project?

12 responses



# Beyond HapMap, what does the world still need?

---

Larger sample sizes

More complete representation of global diversity!

Better discovery and characterization of structural and copy number variation (still an area being developed)

More complete description of the frequency spectrum! Particularly the **low-frequency and rare variants** (minor allele frequency < 0.05)

# The 1000 Genomes Project

---

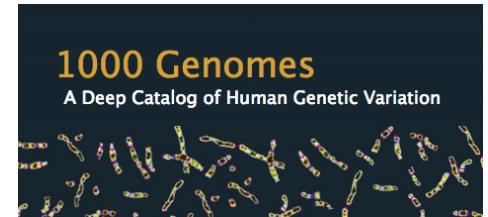
Project ran from 2008-2015

Pilot sequencing started with HapMap3: ten 100kb region by high coverage Sanger Sequencing, 692 individuals in ten HapMap3 populations.

More pilots:

- Two trios (CEU and YRI) at high coverage (42x)
- 179 individuals (CEU, YRI, CHB, JPT) genome-wide at low coverage (2x-6x)
- 8,140 exons from 906 genes (~1.4Mb) at high coverage (>50x) in 697 individuals in seven HapMap3 populations

The International HapMap 3 Consortium, Nature 2010  
The 1000 Genomes Project Consortium, Nature 2010



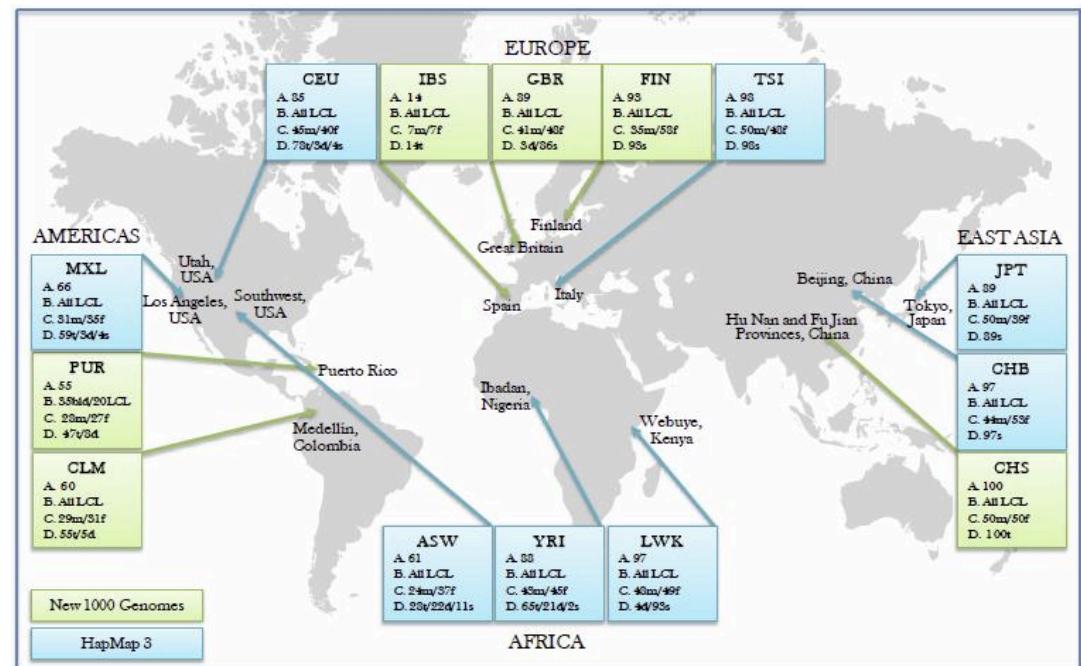
# The 1000 Genomes Project

Phase 1: completed 2012

Sequenced the entire genome of 1,092 individuals from 4 super-populations (EUR, EAS, AFR, AMR; 14 populations total).

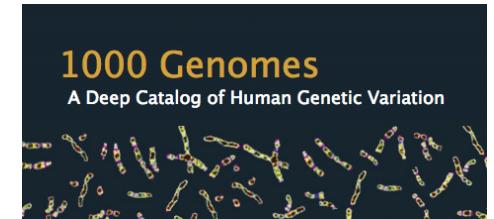
Next-gen sequencing technology (~4x; Illumina, 454, SOLiD...)

38M SNPs, 1.4M indels, estimate captured 98% of accessible SNPs with MAF > 0.01



The 1000 Genomes Project Consortium, Nature 2012

# The 1000 Genomes Project



## Phase 3: completed 2015

2,504 individuals from 5 super-populations (EUR, EAS, AFR, AMR, SAS; 26 populations total).

NGS Illumina only, ~7x

85M SNPs, 3.6M indels. 64M have  
MAF < 0.5%.

Estimate captured 99% of accessible SNPs with MAF > 0.01.



The 1000 Genomes Project Consortium, Nature 2015

# The future?



NHLBI Trans-Omics for Precision Medicine

TOPMed: began in 2014. ~149K individuals from >80 disease cohorts.

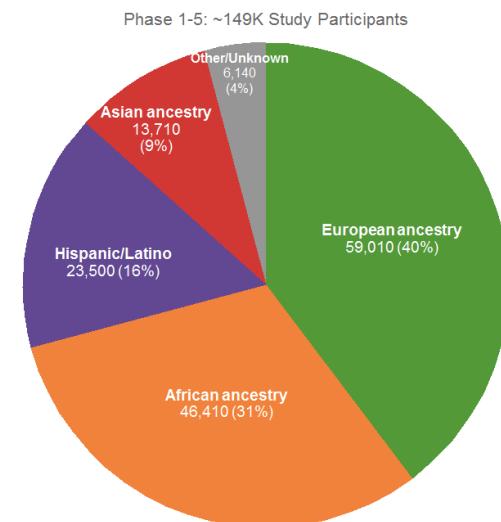
Median depth ~30X whole genome.

Freeze 5 on 53.5K individuals now “available”

- Taliun et al., bioRxiv 2019
- > 400M SNPs and indels, 97% are < 1% MAF, 46% singletons

Freeze 8 on > 140K individuals expected early 2020

- > 800M SNPs, > 66M indels ...



# In ~two decades since HGP...

---

Project	HapMap		
Phase	1	2	3
Year Complete	2005	2007	2010
Sample Size	269	270	1184
Populations	4	4	11
# variants	1 M	3.1 M	1.6 M + CNVs

\* Phase 2 with ~1700 samples sequenced was largely used for methods development

Exome Aggregation Consortium (ExAC), 2016:

60,706 individual exomes (7.1M SNPs, 0.3M indels)

Genome Aggregation Database (gnomAD), 2017-2019:

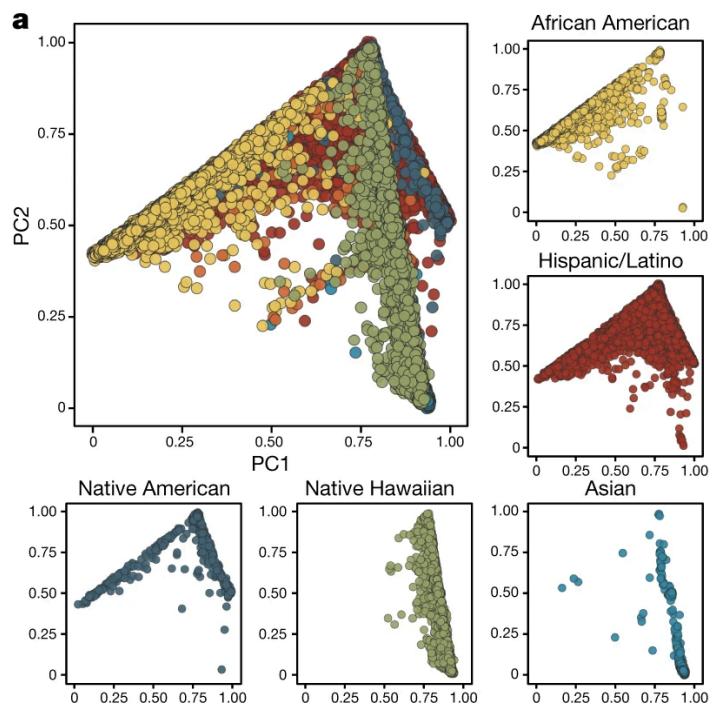
125,748 exomes (16M SNPs, 1.2M indels), 71,702 genomes (602M SNPs, 105M indels)

Also missing, all the major biobanks (UKB, 500K individuals that will be all WGS; BBJ > 160K individuals; etc.)

# Side note about “populations”...

How many populations did TOPMed sequence?

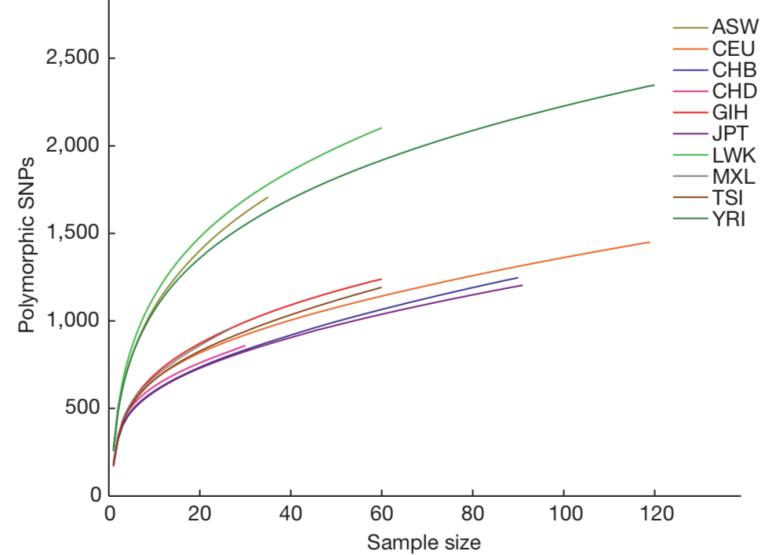
- Sample of ‘convenience’
- Disease cohorts in the United States
- But what is a population anyways?



# What did we learn from these large databases of genetic variation?

---

African populations have more SNPs (per Mb)



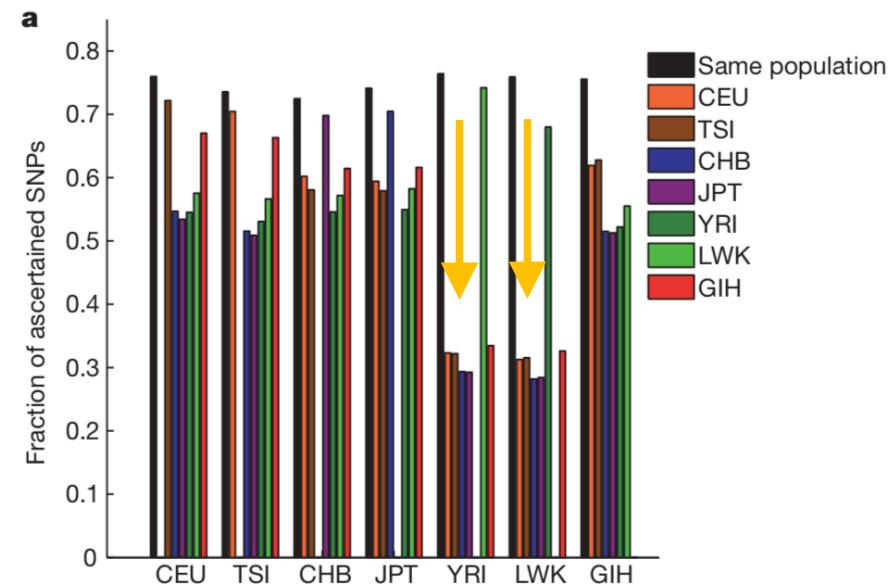
International HapMap3 Consortium, Nature 2010

# What did we learn from these large databases of genetic variation?

African populations have more SNPs (per Mb)

SNPs in non-Africans tend to be a subset of SNPs in Africans

Are SNPs ascertained in 30 individuals from x-axis population polymorphic in another 30 individuals from color-coded population?



International HapMap3 Consortium, Nature 2010

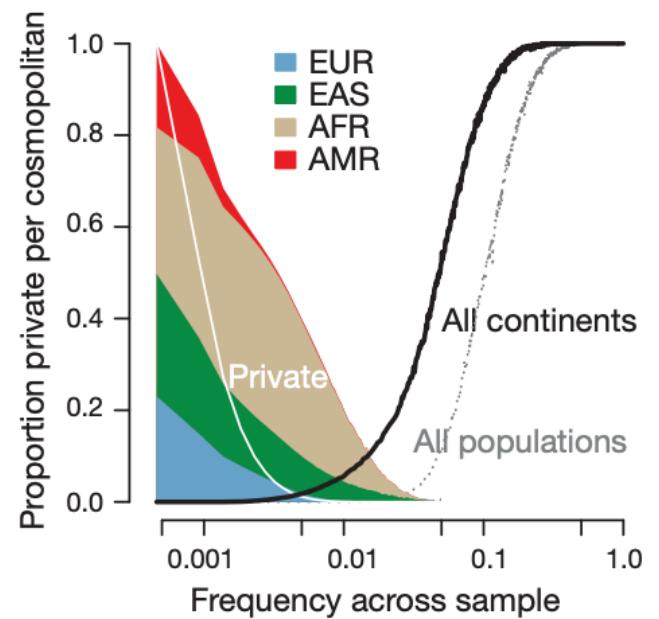
# What did we learn from these large databases of genetic variation?

---

African populations have more SNPs (per Mb)

SNPs in non-Africans tend to be a subset of SNPs in Africans

Common variants are shared across populations, but rare variants are often population-specific



The 1000 Genomes Project Consortium, Nature 2012

# What did we learn from these large databases of genetic variation?

---

African populations have more SNPs (per Mb)

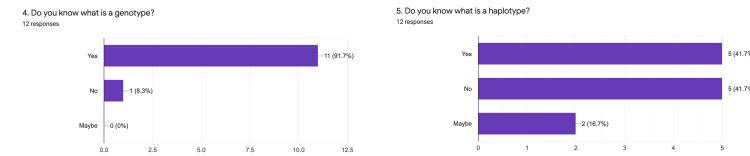
SNPs in non-Africans tend to be a subset of  
SNPs in Africans

Common variants are shared across  
populations, but rare variants are often  
population-specific

The **block-like structure** of the genome

# Some definitions first, just so we're clear

---

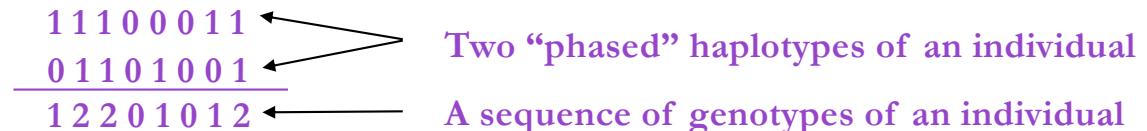


**Haplotype:** description of SNP alleles on a chromosome

- A vector of 0's and 1's, if e.g. 0 denote the ancestral allele, 1 denote the derived allele

**Genotype:** description of alleles on both chromosomes in a person

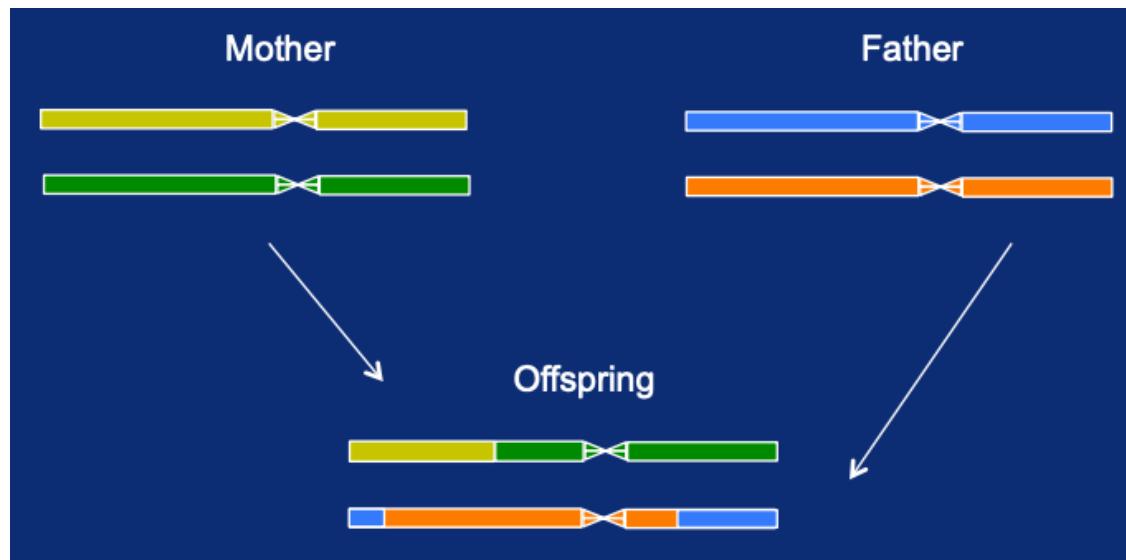
- A vector of 0, 1, and 2's



Biospecimen containing DNA molecules is a mixture of the two copies of your haploid genome, so genotyping (and usually sequencing) tells you only the genotype. The phase of the haplotype can be statistically computed, using programs like EAGLE, SHAPEIT, BEAGLE ...

# Your genome is inherited in blocks due to recombination being relatively rare

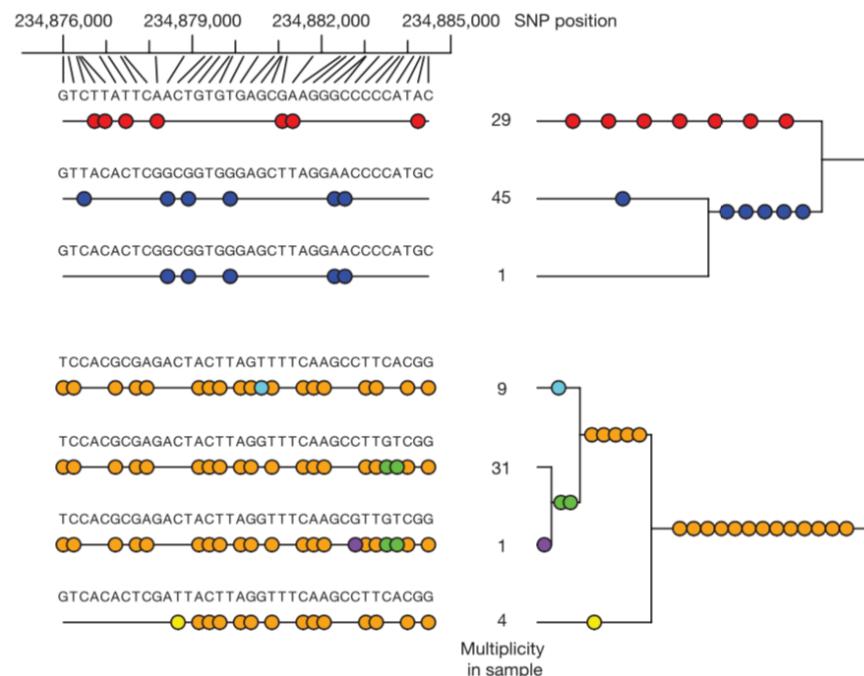
---



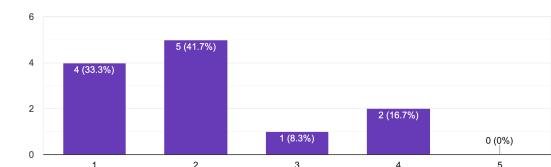
Recombination is uneven across the genome, occurs often in hotspots.

At a e.g. low recombination region of the genome, the entire chunk tend to be inherited together. Thus mutations arising on that chunk over time will be more associated with each other than variation outside of the chunk.

# Block-like structure of the genome



6. How familiar are you with the concept of linkage disequilibrium?  
12 responses



Empirical region on chr2 with 36 SNPs. In theory, that should give rise to  $2^{36}$  different haplotypes. In practice, only 7 were observed (in THIS sample)

The non-random association between two SNPs as a result of this block-like structure, is known as **linkage disequilibrium (LD)**.

HapMap Consortium, Nature 2005

# How is LD measured?

---

Consider two SNPs with frequencies  $p_A$  and  $p_B$  of alleles A and B, with phased data

Main basis is deviation  $D = p_{AB} - p_A p_B$ , but typically we use  $r^2$  or  $D'$

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)}$$

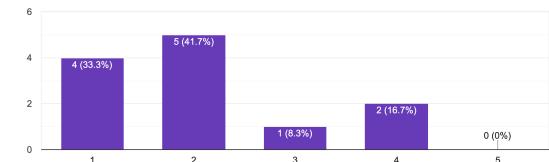
$r^2$  is the correlation coefficient of 1/0 indicator variable indicating the presence of A and B

$$D' = \frac{p_{AB} - p_A p_B}{p_A - p_A p_B}$$

$D'$  has the convenient property that when  $D' = 1$ , it means at least one of the four possible haplotypes is absent.

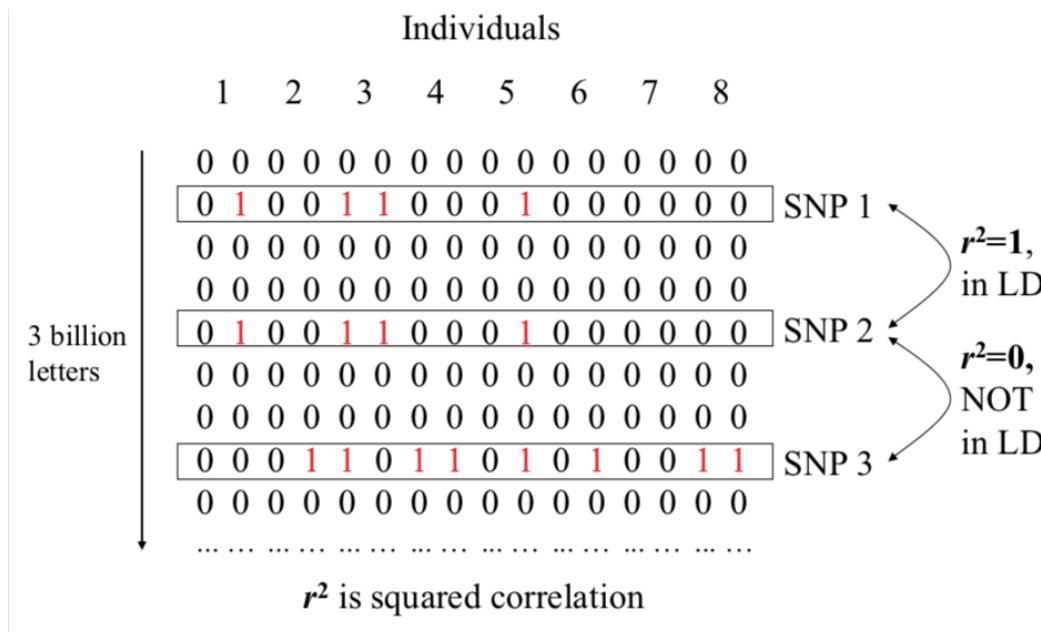
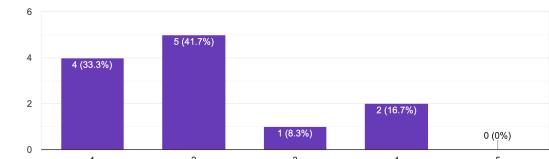
Slatkin, Nat. Rev. Genet. 2008

6. How familiar are you with the concept of linkage disequilibrium?  
12 responses



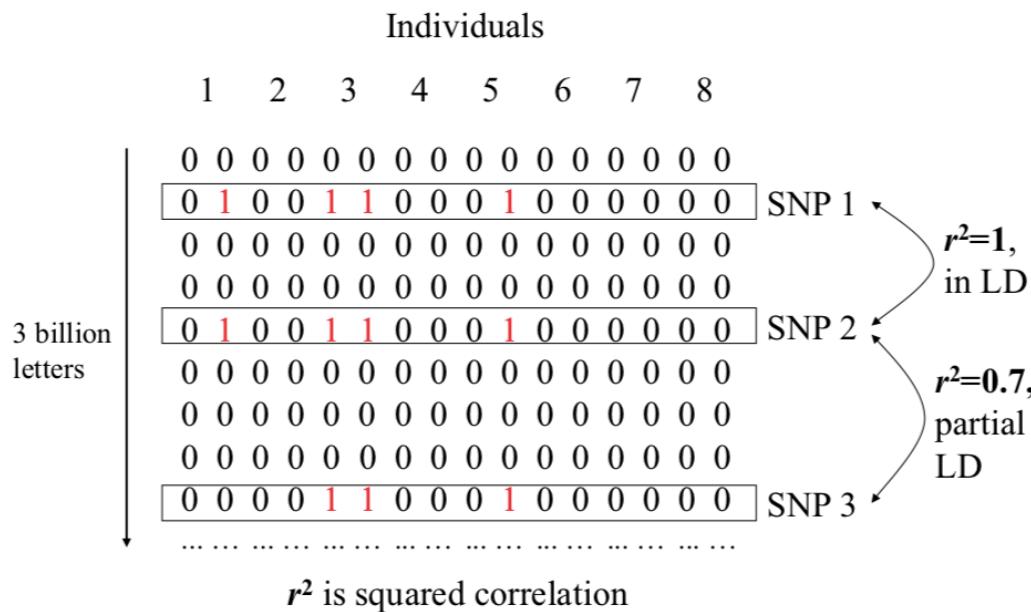
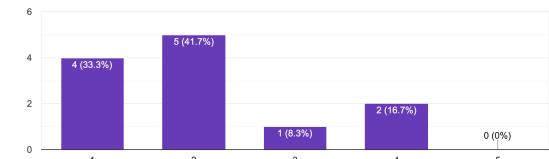
# LD example

6. How familiar are you with the concept of linkage disequilibrium?  
12 responses



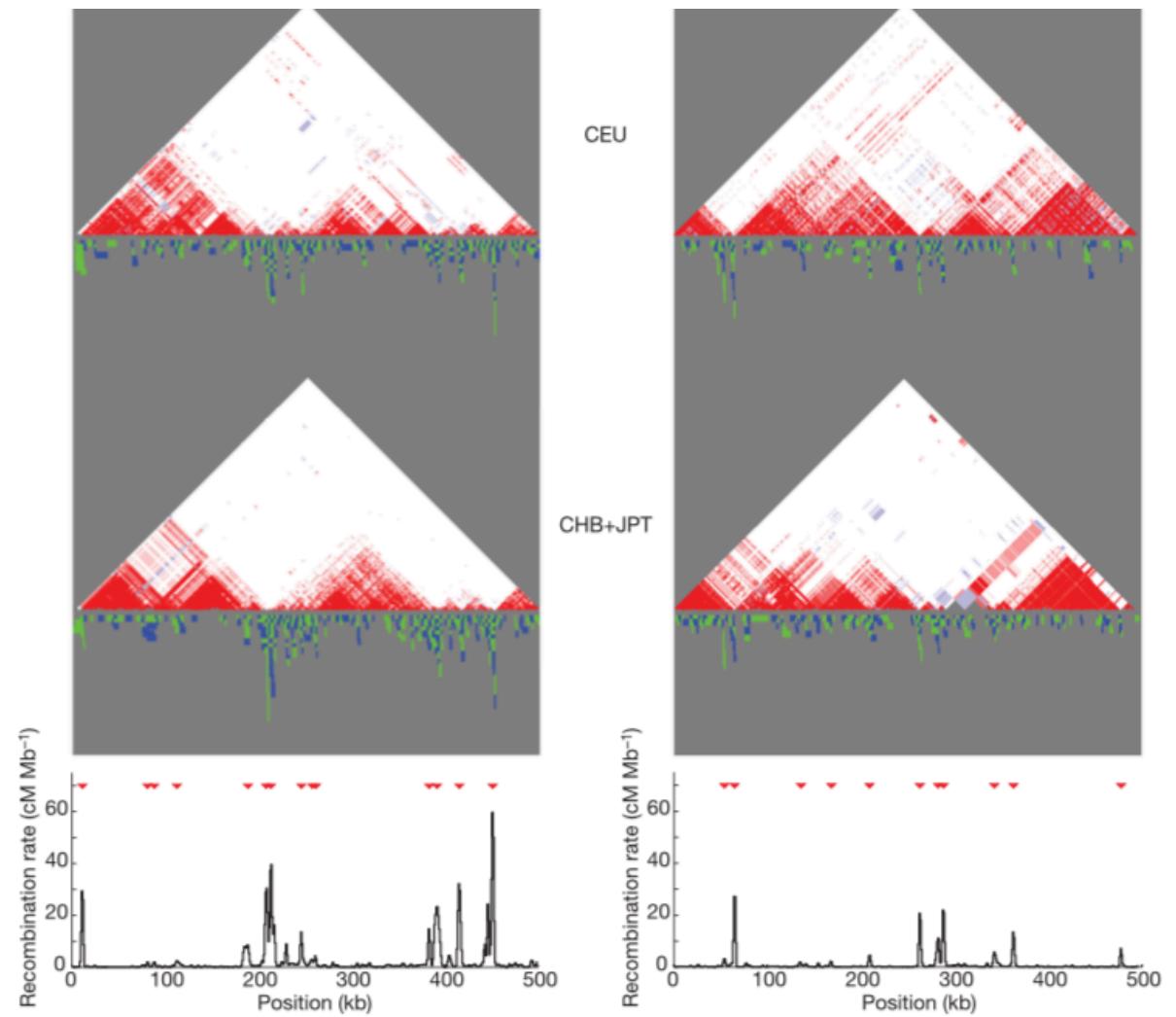
# LD example

6. How familiar are you with the concept of linkage disequilibrium?  
12 responses



# Haplotype blocks

HapMap Consortium, Nature 2005



# Lecture Outline

---

1. ~~Introduction and Motivation~~
2. ~~Pattern of Genetic Variations~~
3. Population Genetic Forces that Impacts Genetic Variations
  - Demography
  - Natural Selection

# Now that we know about genetic variations...

---

Genetic variation is the **difference in DNA** sequences between individuals.

- Mutations contribute to new variations (SNPs if a point mutation)
- Recombination contribute to haplotype diversities (and can be potentially mutagenic, see Halldorsson et al. Science 2019)

The block-like structure enabled Genome-Wide Association Studies (GWAS), because you don't need to genotype and test the *actual* causal allele, but just genotype enough variation in the genome to capture one that is in LD with the causal allele.

**What could have shaped the pattern of variation such that they differ between populations, potentially contribute to differential disease risks between them?**

# Population Genetic forces that shaped genetic variation

---

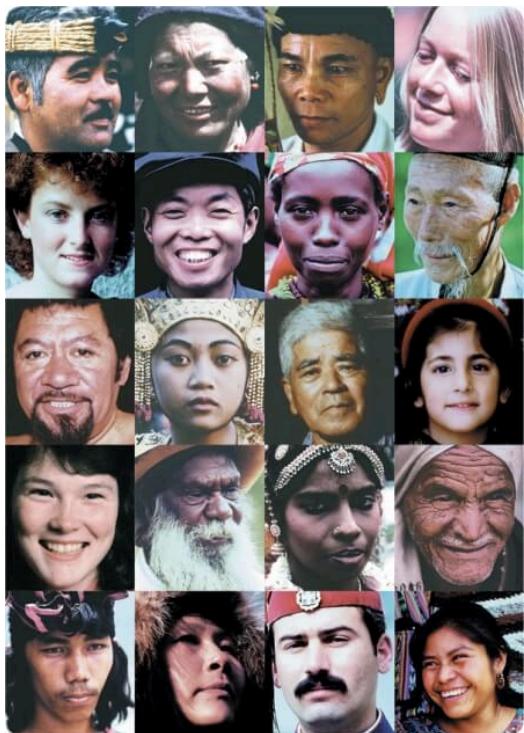
## Demographic history

- Population structure
- Bottleneck
- Admixture

## Natural Selection

# Remember... genetic differences between populations are small, but do exist

---



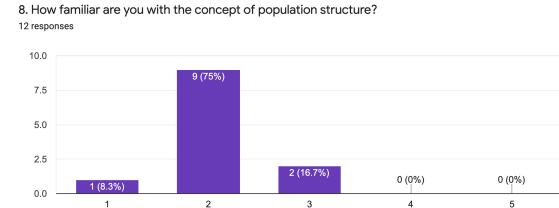
93-95% of human genetic variation is due to variation **within** human populations (Rosenberg et al., Science 2002)

The predicted proportion of observed heterozygosity at a given genetic locus, **assuming human populations are randomly mating**, result in *only* an average error of 5-15% (Novembre & Peter, Curr Opin Genet Dev, 2016)

So, there is fine-scale structure in human populations. A small effect, but existent.

# Population Structure

---



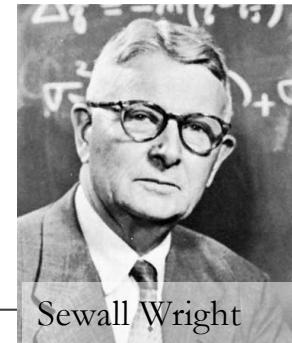
**Population structure** or **population subdivision** refers to genetic differences between (typically discrete) populations due to geographic ancestry

This is due to assortative mating. In other words, mating takes place within sub-groups of the whole population

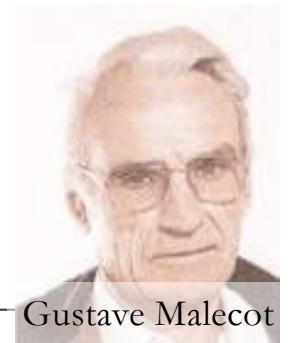
A common model that could lead to population structure in humans is the **Isolation by Distance** model (Wright 1943; Malecot 1948)

Though the effect of population structure is relatively small, it can significantly confound genome-wide association studies of human phenotypes

# Isolation-by-distance



Sewall Wright



Gustave Malecot

**Isolation-by-distance** is a simple consequence of limited dispersal across space. If pairs of populations are close to each other, they will be more genetically similar to each other than populations farther away from each other.

This is not because there is any selective need for genetic similarities, just because individual critters, or their seeds, or pollen, or larvae, etc. are less likely to travel longer distances.

But do note that other models exist, like Kimura's stepping stone model.

<https://www.molecularecologist.com/2012/09/isolating-isolation-by-distance/>

# How to detect/visualize population structure?

---

“Discrete” population clusters

- Model-based clustering (STRUCTURE, FRAPPE, ADMIXTURE, TeraStructure, etc.)

“Continuous” population structure

- Principal Components Analysis (PCA)

# Model-based clustering

---

Example: POP1 and POP2 with known allele frequencies

	SNP1	SNP2	SNP3	SNP4	...	
POP1	0.25	0.57	0.29	0.38	...	(allele frequencies)
POP2	0.40	0.32	0.84	0.22	...	(allele frequencies)
Ind X	2	0	1	1	...	(SNP genotypes)

Does individual X belong to POP1 or POP2?

$P(\text{DATA} \mid X \text{ in POP1})$  is proportional to:

$$(0.25)^2(0.75)^0(0.57)^0(0.43)^2(0.29)^1(0.71)^1(0.38)^1(0.62)^1 = 0.0006$$

$P(\text{DATA} \mid X \text{ in POP2})$  is proportional to:

$$(0.40)^2(0.60)^0(0.32)^0(0.68)^2(0.84)^1(0.16)^1(0.22)^1(0.78)^1 = \mathbf{0.0017}$$

# Model-based clustering

---

Example: POP1 and POP2 with known allele frequencies

	SNP1	SNP2	SNP3	SNP4	...	
POP1	0.25	0.57	0.29	0.38	...	(allele frequencies)
POP2	0.40	0.32	0.84	0.22	...	(allele frequencies)
Ind X	2	0	1	1	...	(SNP genotypes)

If individual X has ancestry  $\alpha$  from POP1 and  $(1 - \alpha)$  from POP2, then what is the most likely value of  $\alpha$ ?

$P(\text{DATA} | \alpha)$  is proportional to:

$$[0.25\alpha + 0.40(1-\alpha)]^2 [0.75\alpha + 0.60(1-\alpha)]^0 [0.57\alpha + 0.32(1-\alpha)]^0 [0.43\alpha + 0.68(1-\alpha)]^2 ] \text{ max value } 0.0020 \\ [0.29\alpha + 0.84(1-\alpha)]^1 [0.71\alpha + 0.16(1-\alpha)]^1 [0.38\alpha + 0.22(1-\alpha)]^1 [0.62\alpha + 0.78(1-\alpha)]^1 ] \text{ attained at } \alpha = 0.22$$

# Model-based clustering

---

These ideas can be generalized to  $M$  SNPs,  $N$  populations, with known allele frequencies  $p_{mn}$  for SNP  $m$  in population  $n$ , observed genotype count  $g_m$  for SNP  $m$  in individual X. Then one can assign the population  $n = 1$  to  $N$  that individual belongs to, or assign the most likely fractional membership  $\alpha$  to each population  $n$ .

# Model-based clustering

---

This can be further generalized to many individuals  $X_i$ , with unknown allele frequency  $p_{mn}$  for SNP  $m$  in population  $n$

# Model-based clustering

---

This can be further generalized to many individuals  $X_i$ , with unknown allele frequency  $p_{mn}$  for SNP  $m$  in population  $n$

Then you model the joint likelihood:

$P(\text{DATA} \mid X_i \sim \alpha_{i1}, \dots, \alpha_{iN} \text{ for each } i; p_{mn})$  is proportional to

$$\prod_{i=1}^I \prod_{m=1}^M (\sum_{n=1}^N \alpha_{in} p_{mn})^{g_{im}} (\sum_{n=1}^N \alpha_{in} (1 - p_{mn}))^{2-g_{im}}$$

Then find values of  $\alpha_{in}, p_{mn}$  which maximize this likelihood.

# Model-based clustering

---

This can be further generalized to many individuals  $X_i$ , with unknown allele frequency  $p_{mn}$  for SNP  $m$  in population  $n$

Then you model the joint likelihood:

$P(\text{DATA} \mid X_i \sim \alpha_{i1}, \dots, \alpha_{iN} \text{ for each } i; p_{mn})$  is proportional to

$$\prod_{i=1}^I \prod_{m=1}^M (\sum_{n=1}^N \alpha_{in} p_{mn})^{g_{im}} (\sum_{n=1}^N \alpha_{in} (1 - p_{mn}))^{2-g_{im}}$$

Then find values of  $\alpha_{in}, p_{mn}$  which maximize this likelihood.

The different approaches to maximize this likelihood is adopted by different program (EM, MCMC, variational Bayes approximations, etc.)

# Model-based clustering

---

STRUCTURE (Pritchard et al. Genetics 2000)

FRAPPE (Tang et al. Genet. Epidemiol. 2005)

ADMIXTURE (Alexander et al. Genome Research 2009)

TeraStructure (Gopalanan et al. Nat. Genetics 2016)

...

# Model-based clustering in action

---

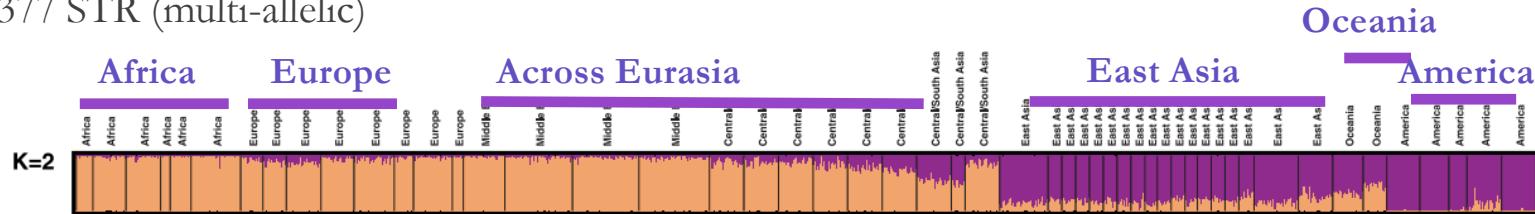
Human Genome Diversity Panel data

- 1,056 individuals, 52 world populations
- 377 STR (multi-allelic)

# Model-based clustering in action

Human Genome Diversity Panel data

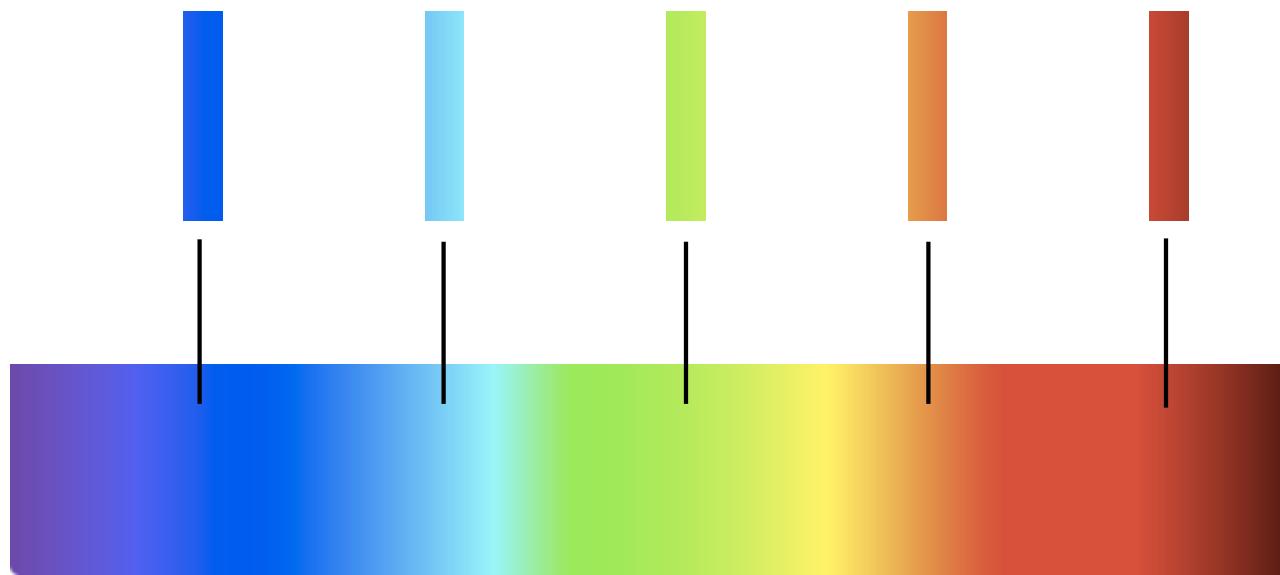
- 1,056 individuals, 52 world populations
- 377 STR (multi-allelic)



Rosenberg et al. Science 2002

# Does “race” exist? Not by genetics

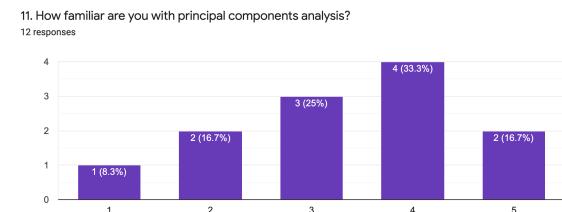
---



# Principal Components Analysis

---

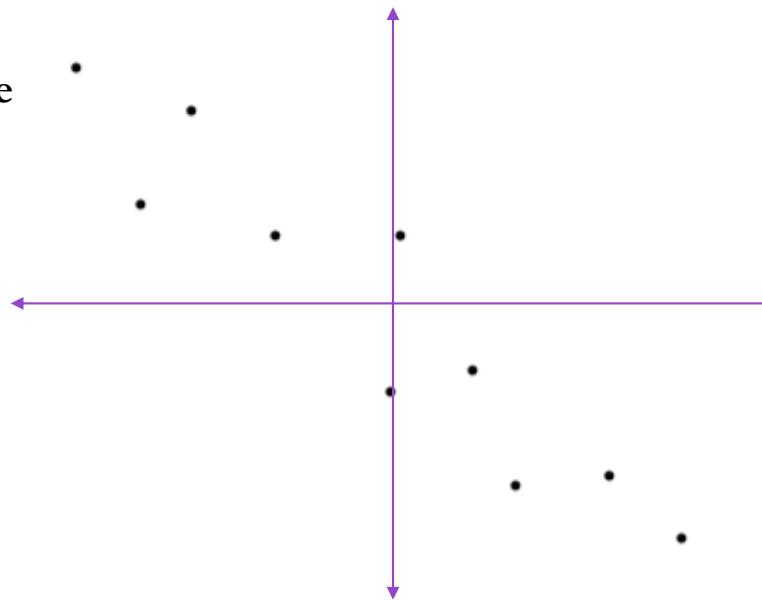
**Principal Components Analysis**, or **PCA**, is a dimension reduction technique that converts high dimensional data (such as human SNP data) into a set of linearly uncorrelated variables (principal components, or PCs). The transformation is done in such a way that the first principal component has the largest possible variance explained.



# PCA, an intuition

---

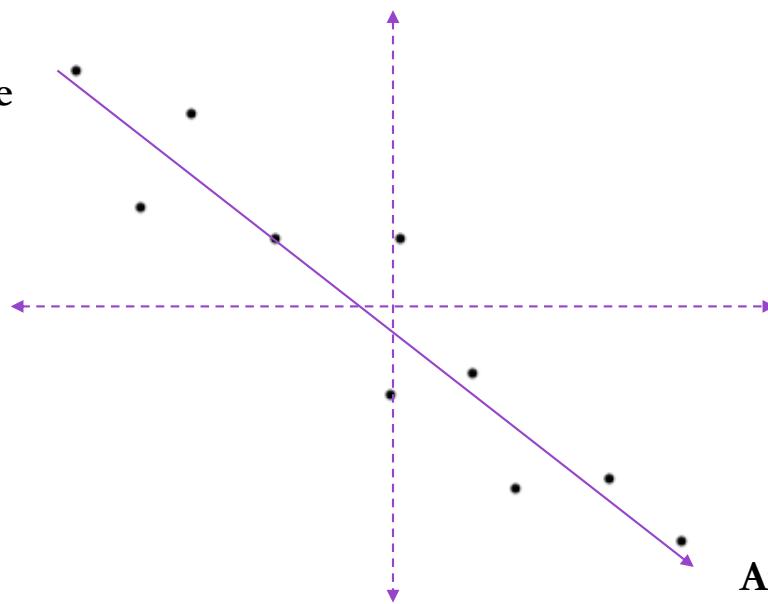
10 points in 2-dimensional space



# PCA, an intuition

---

10 points in 2-dimensional space

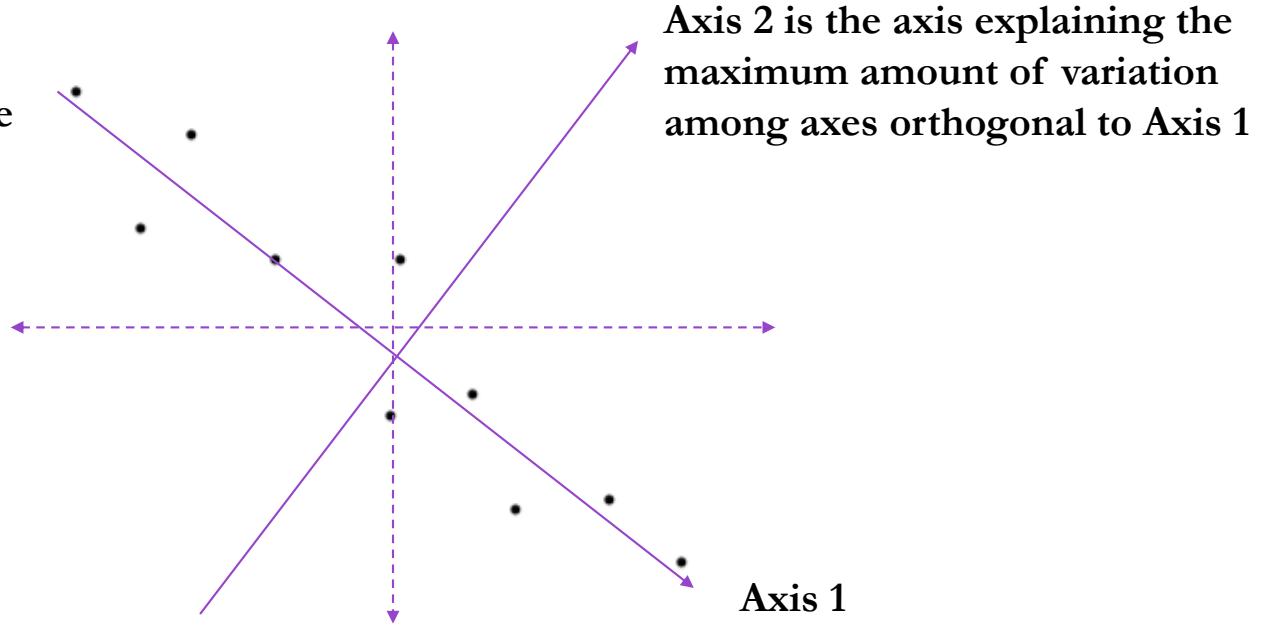


Axis 1 is the axis explaining the max amount of variation

# PCA, an intuition

---

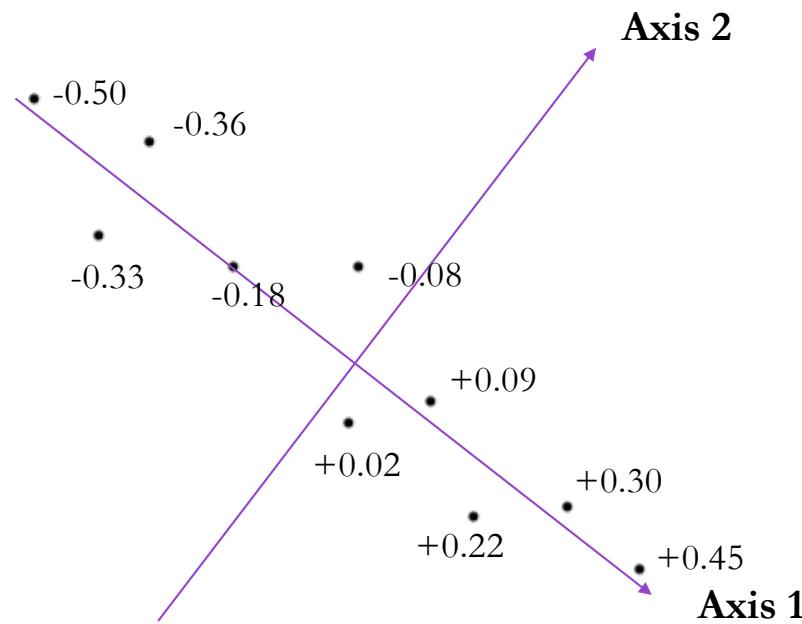
10 points in 2-dimensional space



# PCA, an intuition

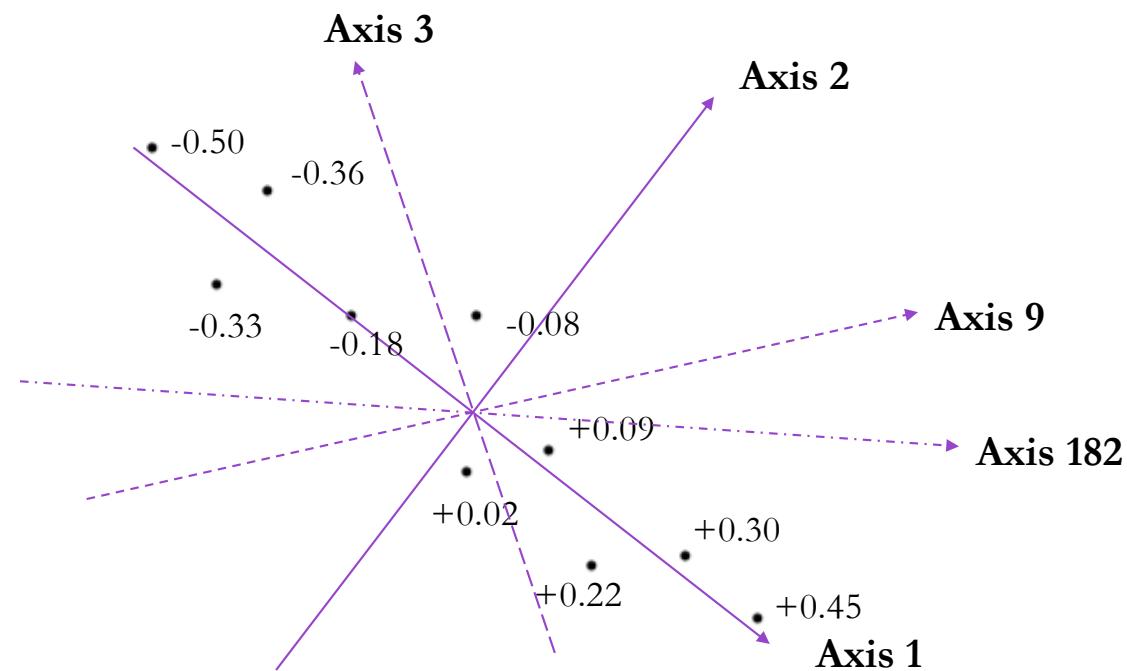
---

10 points in 2-dimensional space



# PCA, an intuition

10 points in 10000-D space



# Principal Components Analysis

---

When applied to genetic data, it can be used to explain differences among individuals. The top PCs are viewed as continuous axes of variation that reflect genetic variation due to (usually) geographical ancestry in the sample.

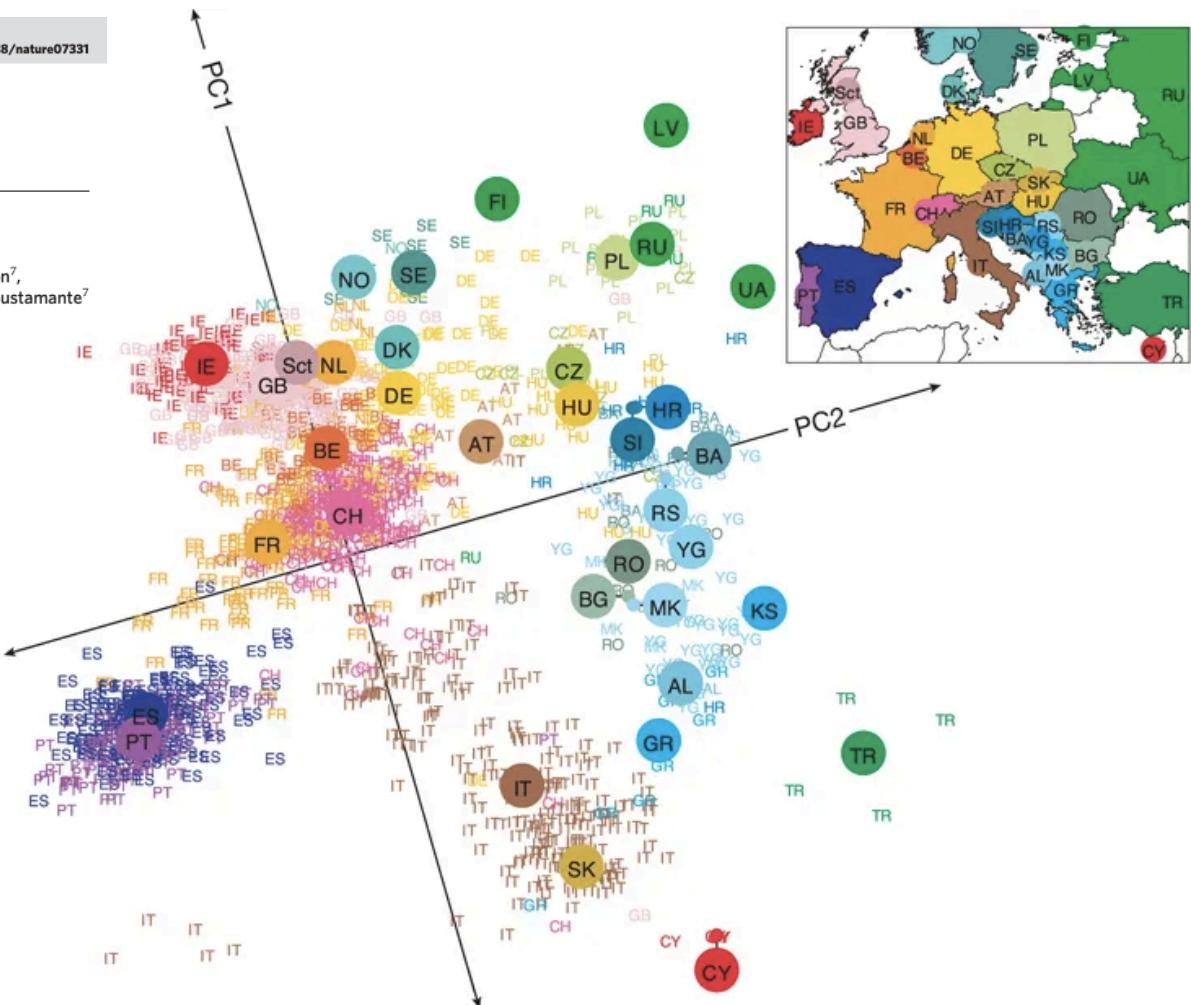
Individuals with similar values for a particular top PC will have similar ancestry for that axes.

## LETTERS

## **Genes mirror geography within Europe**

John November<sup>1,2</sup>, Toby Johnson<sup>4,5,6</sup>, Katarzyna Bryc<sup>7</sup>, Zoltán Kutalik<sup>4,6</sup>, Adam R. Boyko<sup>7</sup>, Adam Auton<sup>7</sup>, Amit Indap<sup>7</sup>, Karen S. King<sup>8</sup>, Sven Bergmann<sup>4,6</sup>, Matthew R. Nelson<sup>8</sup>, Matthew Stephens<sup>2,3</sup> & Carlos D. Bustamante<sup>7</sup>

1,387 Europeans in POPRES dataset  
Affymetrix 500K array



## A Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group—Han Chinese

Charleston W.K. Chiang,<sup>\*1,2</sup> Serghei Mangul,<sup>3,4</sup> Christopher Robles,<sup>5</sup> and Sriram Sankararaman<sup>3,5</sup>

<sup>1</sup>Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA

<sup>2</sup>Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA

<sup>3</sup>Department of Computer Science, University of California Los Angeles, Los Angeles, CA

<sup>4</sup>Institute for Quantitative and Computational Bioscience, University of California Los Angeles, Los Angeles, CA

<sup>5</sup>Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA

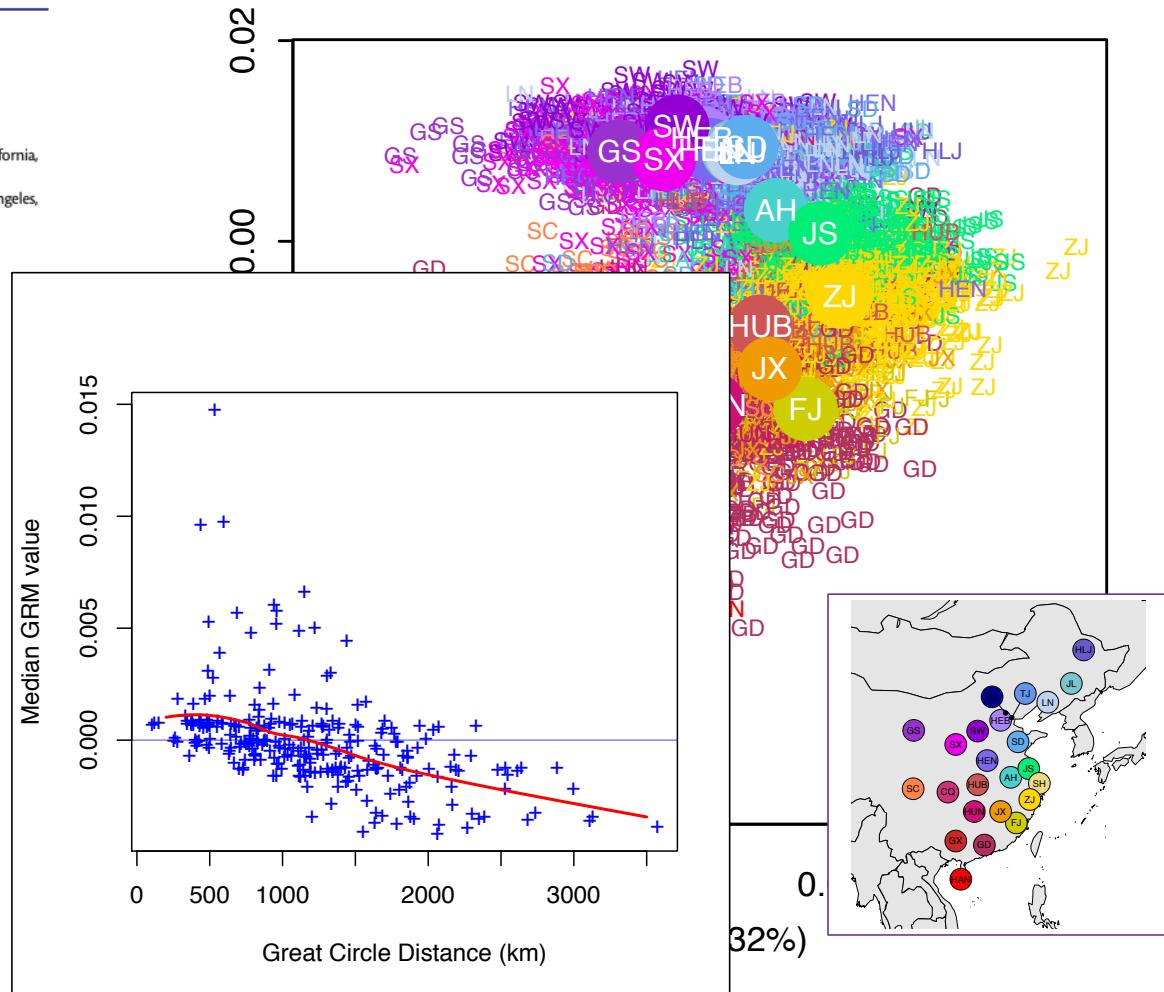
\*Corresponding author: E-mail: charleston.chiang@med.usc.edu.

Associate editor: Connie Mulligan

7,457 Han Chinese from 19/22 provinces  
1-2x low coverage WGS data

Continuous cline corresponding to geography  
PC1 correlated with latitude ( $r = 0.88$ )  
PC2 correlated with longitude ( $r = 0.70$ )

Consistent with isolation-by-distance:  
Populations closer to each other geographically  
have higher genetic similarities



# PCs do not necessarily reflect geography, or even population structure ...

---

Batch effects (see Clayton et al. Nat. Genet. 2005; Price et al. Nat. Genet. 2006)

Cryptic relatedness (see Patterson et al. 2006 PLoS Genet.)

Long-range LD, e.g. due to inversion polymorphisms (see Tian et al. 2008 PLoS Genet., Price et al. AJHG 2008)

# Really quick word on haplotype-based clustering

---

Allele frequency-based approach ignores linkage information and treating each marker in analysis as independent (usually pruning to quasi-independent subset). Haplotypes can be more informative for clustering.

E.g. chromopainter and fineSTRUCTURE (Lawson et al. PLoS Genet. 2012)

Down-side is that usually haplotype-based approach is computationally intensive, thus not scalable to the sizes of datasets these days. Works best for < 2,000 individuals.

- But this is active area of development, with promise! (e.g. pBWT, Durbin, Bioinformatics 2014)

# Population Genetic forces that shaped genetic variation

---

Demographic history

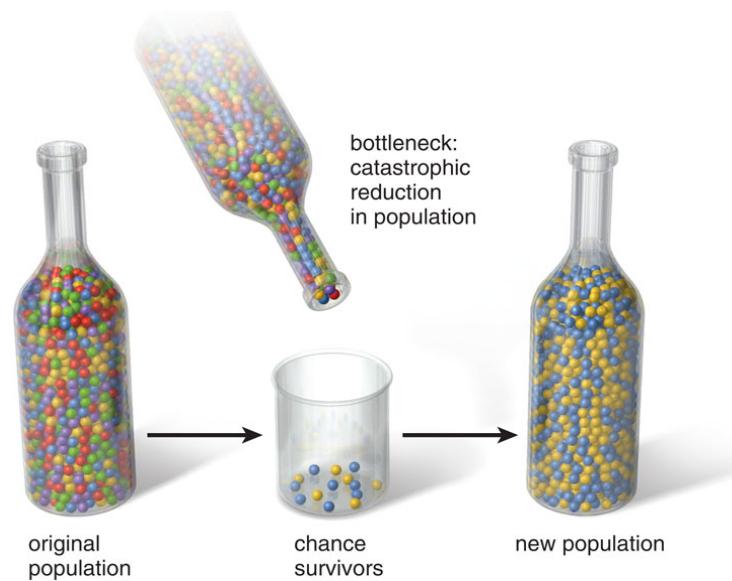
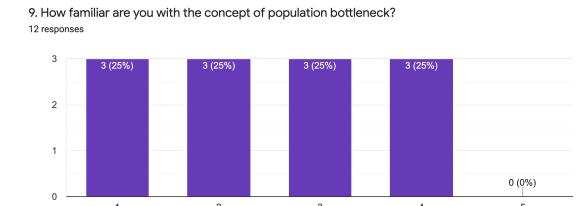
◦ Population structure

- Bottleneck
- Admixture

Natural Selection

# Population bottleneck

**Population bottleneck** is an event that drastically reduces the size of a population

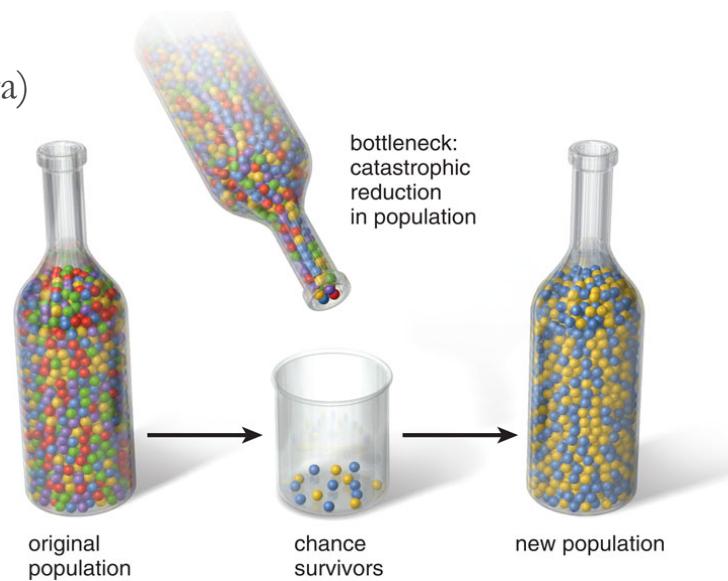


# Population bottleneck

**Population bottleneck** is an event that drastically reduces the size of a population

It could occur by...

- “Environmental” disasters (Native Americans during colonial era)
- Founding event of a new population (Finland)

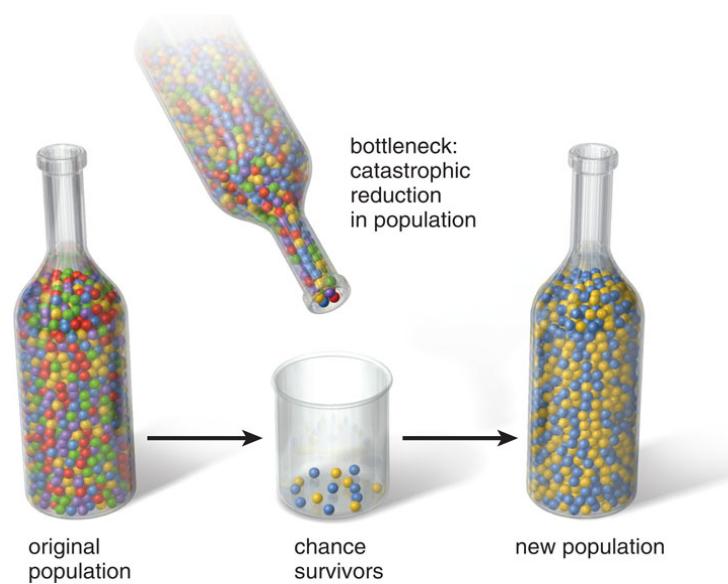


# What are the impact due to bottleneck?

---

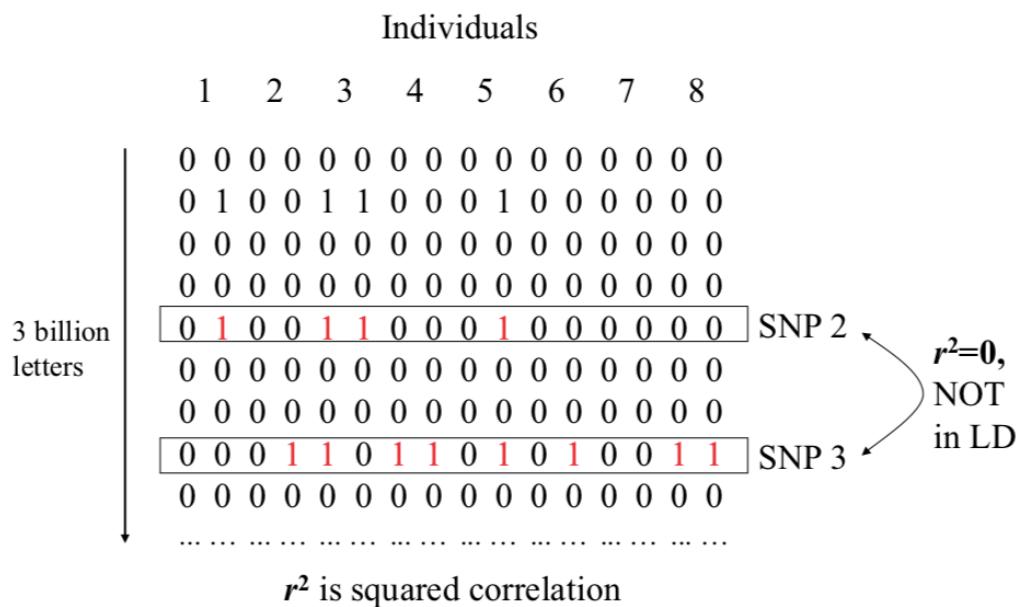
Population bottleneck decreases the diversity of the gene pool, because many alleles in the parental populations would be lost.

Because of limited number of surviving lineages through a bottleneck, there will be increased LD (e.g. out-of-Africa event, founding of Finland).



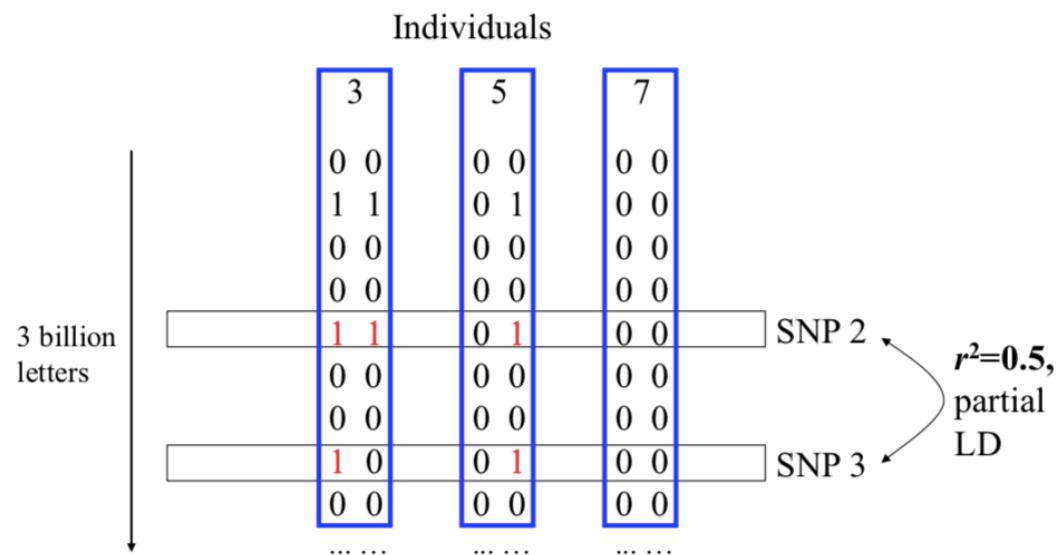
# What are the impact due to bottleneck?

---



# What are the impact due to bottleneck?

# What are the impact due to bottleneck?



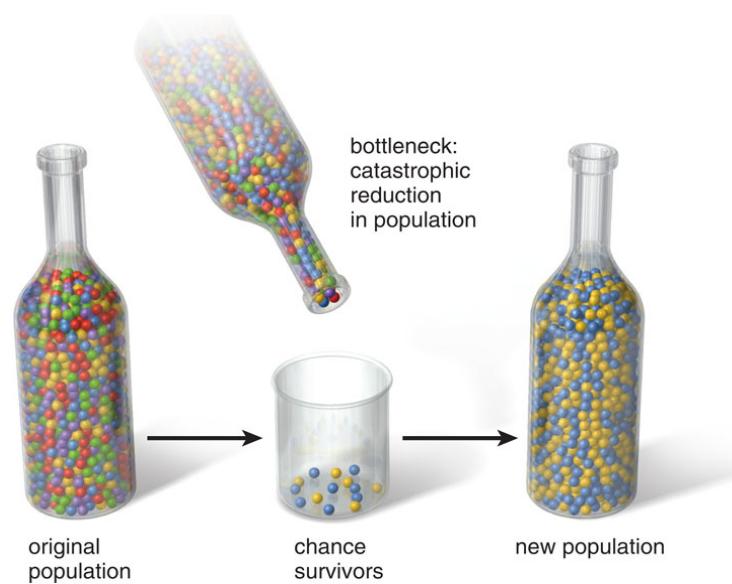
# What are the impact due to bottleneck?

---

Population bottleneck decreases the diversity of the gene pool, because many alleles in the parental populations would be lost.

Because of limited number of surviving lineages through a bottleneck, there will be increased LD (e.g. out-of-Africa event, founding of Finland).

Because of small population sizes, the impact of genetic drift\* would increase, causing some rare alleles to elevate in frequency by chance.

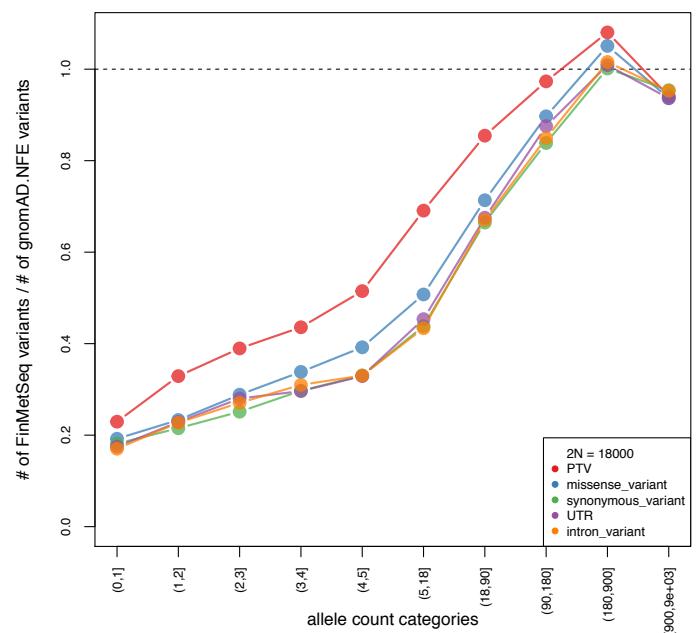


# Enrichment of (deleterious) alleles

Because of small population sizes, the impact of **genetic drift\*** would increase, causing some rare alleles to elevate in frequency by chance.

- Genetic drift is the random fluctuation of frequency of alleles due to sampling alleles across generations.
- In a smaller population, the sampling error is larger, hence frequencies can change more drastically.

Compared to a larger population, the enrichment would be more pronounced for deleterious alleles that would otherwise be negatively selected against.



Locke\*, Steinberg\*, Chiang\*, Service\* et al., Nature 2019

# How to model the size history?

---

# First, why is it important to model the population size history?

---

Of interest in their own right, for historical or anthropological reasons.

Establishes an important null model for any test of neutrality (Nielsen et al., Genome Res. 2005), disease associations (Mathieson and McVean, Nat. Genet. 2012), or recombination rate inference (Johnston and Cutler, AJHG 2012)

Important for conservation genetics, informing breeding strategies for maintaining genetic diversity in endangered population (Mays et al. Curr. Biol. 2018)

Quite difficult to do for complicated models (multiple populations, migrations, admixtures, size changes, etc.) both statistically and computationally. So tend to assume a simplifying parametric model with a few parameters, or a nonparametric model of a single population.

Spence et al. Curr Op Genet Dev 2018

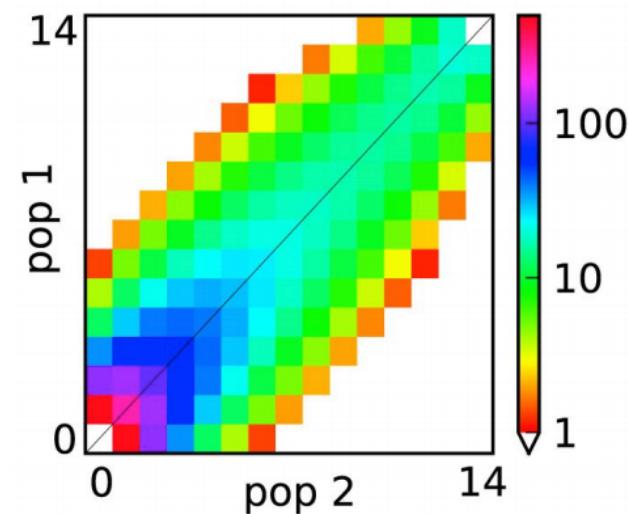
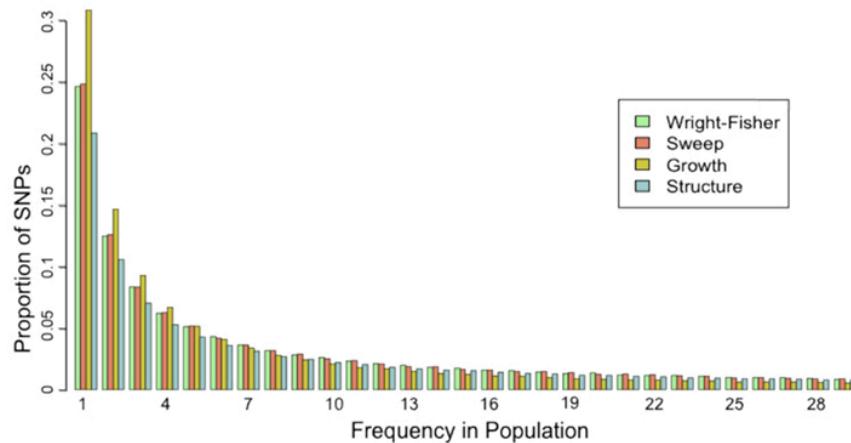
# How to model the size history?

---

Allele frequency-based methods using the multi-population site frequency spectrum (SFS) to infer parametric models.

# Site Frequency Spectrum (SFS)

**Site Frequency Spectrum**, or **SFS**, is a histogram of the frequencies of SNPs in a sample of individuals from a population.



# How to model the size history?

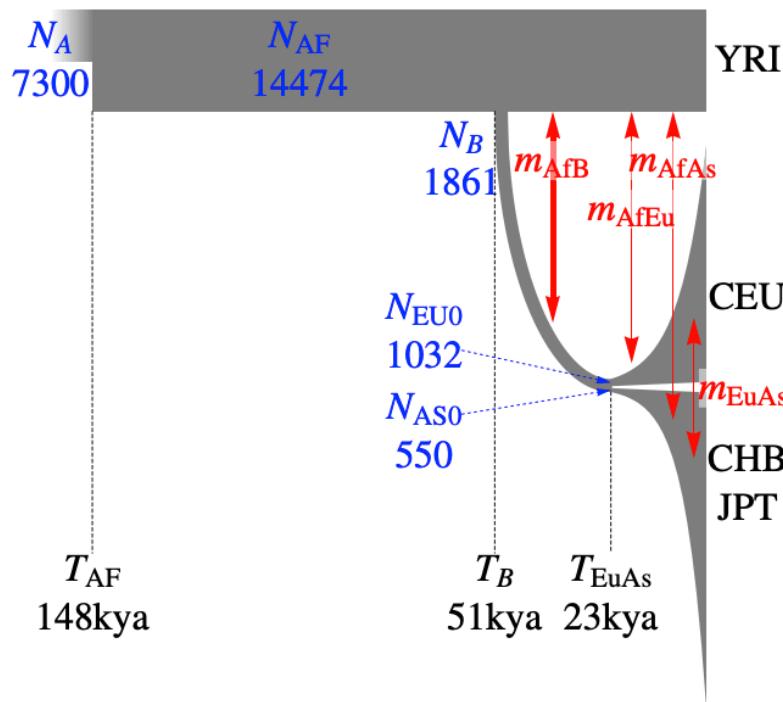
---

Allele frequency-based methods using the multi-population site frequency spectrum (SFS) to infer parametric models.

- Very fast, can explore a bit more complicated model with a few more parameters.
- Assumes all loci are independent, ignore information contained in linkage.
- References: Gutenkunst et al. PLoS Genet. 2009, Gravel et al. PNAS 2012, Excoffier et al. PLoS Genet. 2013, Bhaskar et al. Genome Res. 2015 ...

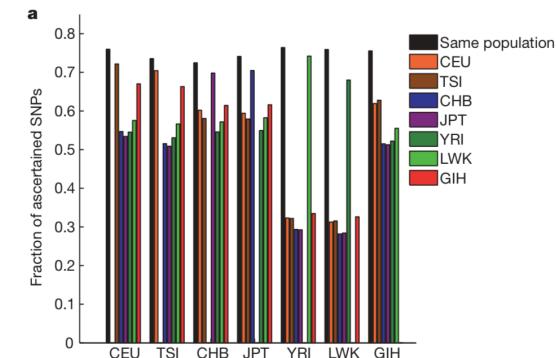
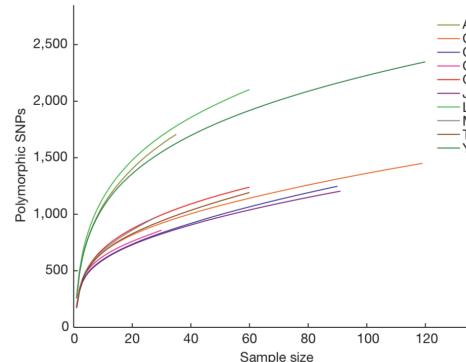
Generally, given a parametric demographic model, multi-dimensional SFS can be simulated (sometimes with approximations) and fitted to the observed data. Find the set of parameters that maximize the (composite) likelihood.

# Demographic history of major human continental populations



Gravel et al. PNAS 2012

Used 1000 Genomes phase 1 data of the 4 populations



The International HapMap3 Consortium, Nature 2010

# How to model the size history?

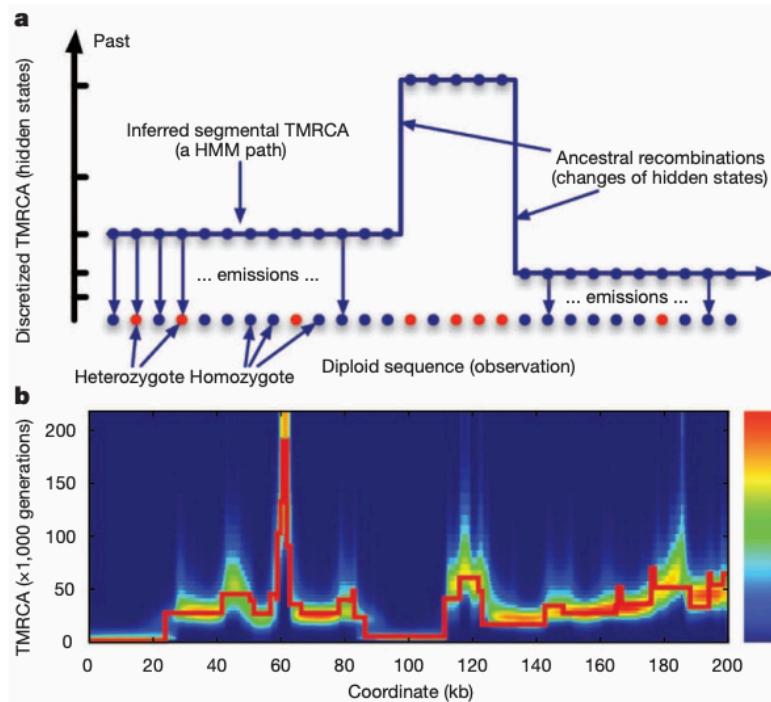
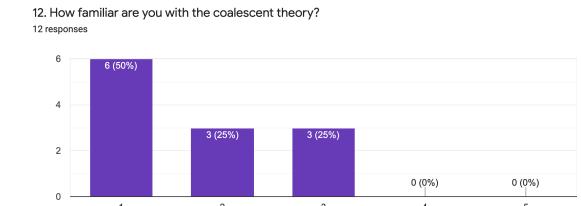
---

Allele frequency-based methods using the multi-population site frequency spectrum (SFS) to infer parametric models.

Haplotype-based methods

- Coalescent HMM
- Identity-by-descent sharing patterns ← won't discuss this today

# Pairwise sequential Markovian coalescent (PSMC)



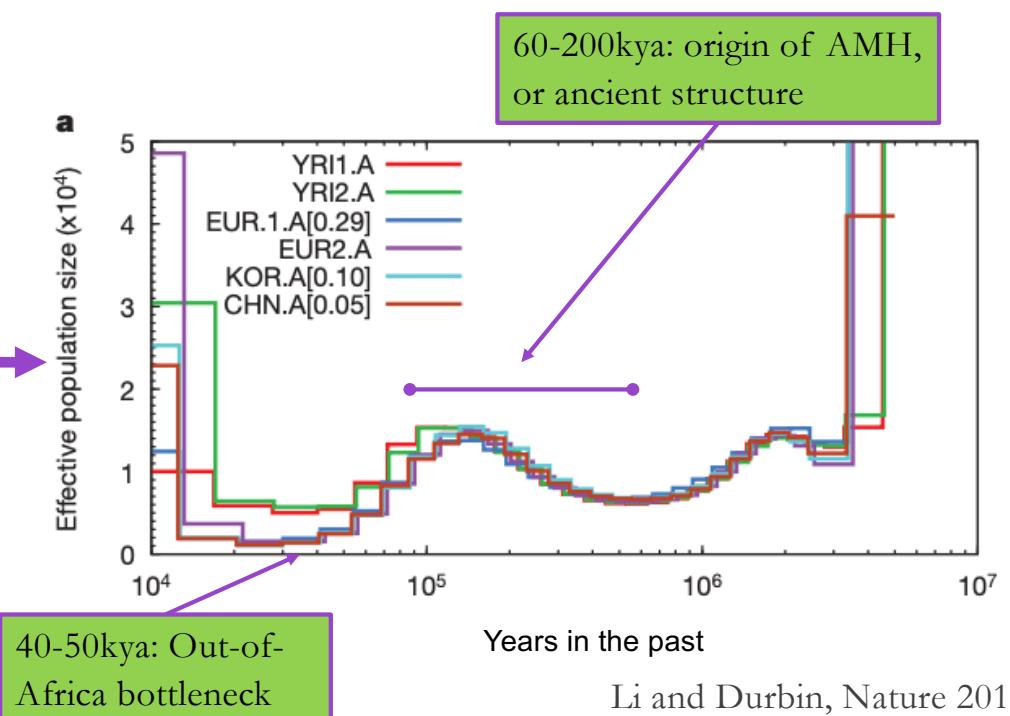
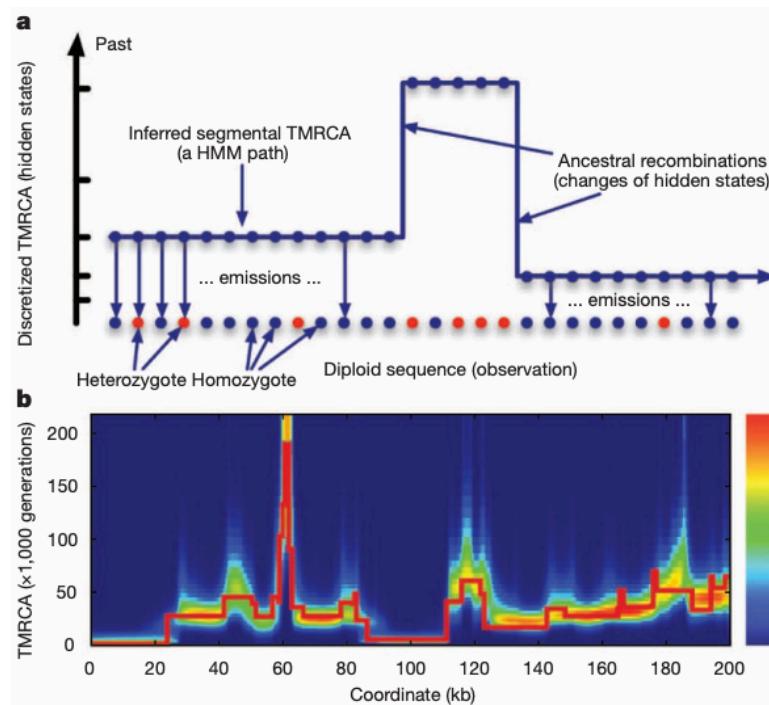
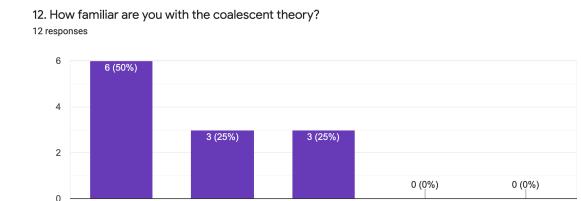
Li and Durbin, Nature 2011

Takes  $N = 1$  as input (2 unphased haplotypes)

Discretize time as hidden state, an HMM to estimate the distribution of time to most common recent ancestor (TMRCA) across the thousands of independent sites in a genome.

The rate of coalescent events is inversely proportional to effective population size ( $N_e$ ). So if at a particular time period there are a lot of loci coalescing, that means  $N_e$  is small during that period.

# Pairwise sequential Markovian coalescent (PSMC)



# Some limitations of PSMC

---

Doesn't recover sudden changes in  $N_e$

Loses power to detect recent size changes (too few coalescent events with just a single diploid genome)

Requires knowing the right mutation rate and generation time to scale coalescent time to actual time.

Changes in  $N_e$  could be due to population structure

Advancements (mostly to increase sample size): MSMC (2014), SMC++ (2017), ASMC (2018) ...

<https://www.molecularecologist.com/2016/05/opening-pandoras-box-psmc-and-population-structure/>

# Population Genetic forces that shaped genetic variation

---

Demographic history

- Population structure
- Bottleneck
- Admixture

Natural Selection

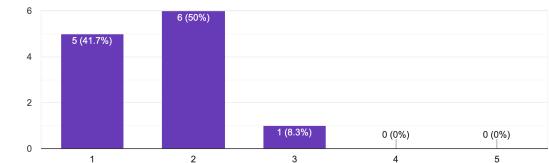
# What is an admixed population?

---

An **admixed population** is a population with “recent” ancestry from two or more continents.

- let’s say, ~ within the last 1000 years, loosely defined.
- Could be used to refer to “ancient admixture”, or “archaic admixture”, like with Neandertals

10. How familiar are you with the concept of population admixture?  
12 responses



# What is the difference between population structure and population admixture?

---

**Structure** is genetic differences due to geographic ancestry. We are usually interested to use genome-wide data to infer broader scale cluster membership.

**Admixture** is mixed ancestry from multiple continental populations. We are usually interested to infer local ancestry at each location in the genome.

Population admixture implies population structure; population structure does not imply population admixture.

# What are some examples of admixed populations (in U.S. or around the world)?

---

African Americans

- African and European ancestry; > 10% of U.S. population



# What are some examples of admixed populations (in U.S. or around the world)?

---

## African Americans

- African and European ancestry; > 10% of U.S. population

## Latino Americans

- European, Native American, and African ancestry; >15% of U.S. population
- e.g. Mexican Americans, Puerto Ricans, etc.
- Hundreds of millions of people throughout Latin America



# What are some examples of admixed populations (in U.S. or around the world)?

---

## African Americans

- African and European ancestry; > 10% of U.S. population

## Latino Americans

- European, Native American, and African ancestry; > 15% of U.S. population
- e.g. Mexican Americans, Puerto Ricans, etc.
- Hundreds of millions of people throughout Latin America

## Native Hawaiians

- Polynesian, European, and East Asian ancestry

## Uyghurs

- East Asian and European-related ancestry

Non-Hispanic Whites is currently the majority in U.S. (77.5% in 2014). They are projected to be the only 43.6% in 2060, while multi-race individuals will grow by 219%.

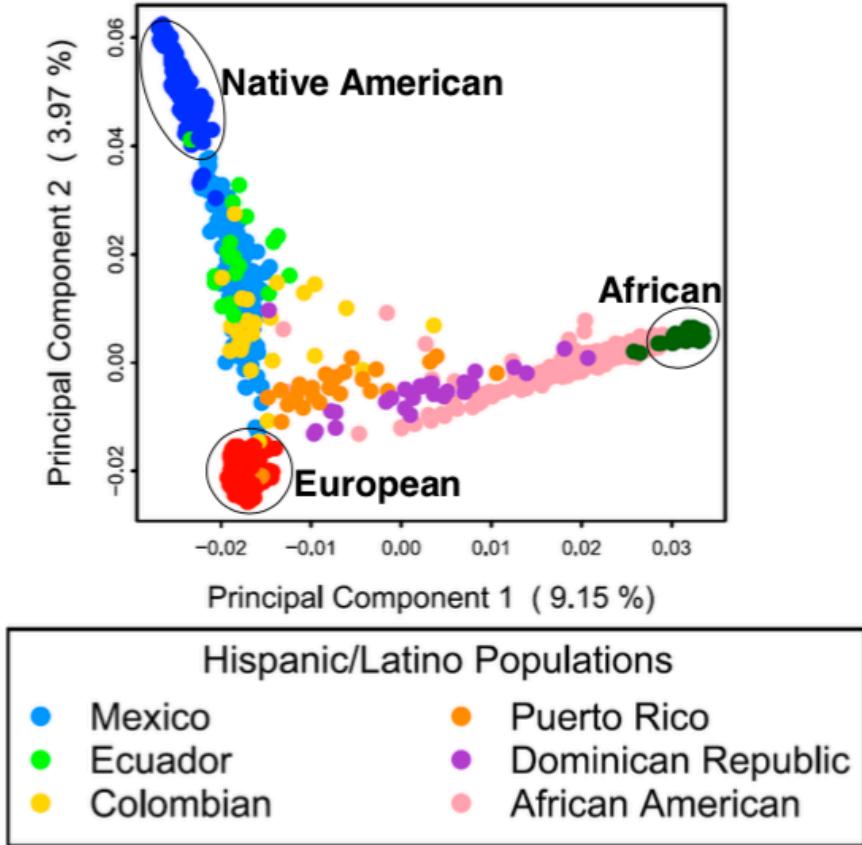
-- U.S. Census Bureau

# How to infer genomic ancestry?

---

Globally

- Apply the clustering programs that allow fractional ancestry (STRUCTURE, FRAPPE, ADMIXTURE, etc.)
- Or, apply PCA. Admixed individuals tend to form a cline between parental populations in PCA space.



	European	Native Am	African
Mexican Americans	~50%	~45%	~5%
Puerto Ricans	~60%	~20%	~20%
Brazilians and Columbians	~70%	~20%	~10%

Within a population, there are substantial variations of the ancestry proportions per individual (the “cline”)

Across geographical space (such as U.S.), there are variation of the average ancestry proportion as well.

Estimates are for population sampled and defined.  
Values may not apply to all populations.

Bryc et al. PNAS 2010

Price et al. AJHG 2007

# How to infer genomic ancestry?

---

## Globally

- Apply the clustering programs that allow fractional ancestry (STRUCTURE, FRAPPE, ADMIXTURE, etc.)
- Or, apply PCA. Admixed individuals tend to form a cline between parental populations in PCA space.

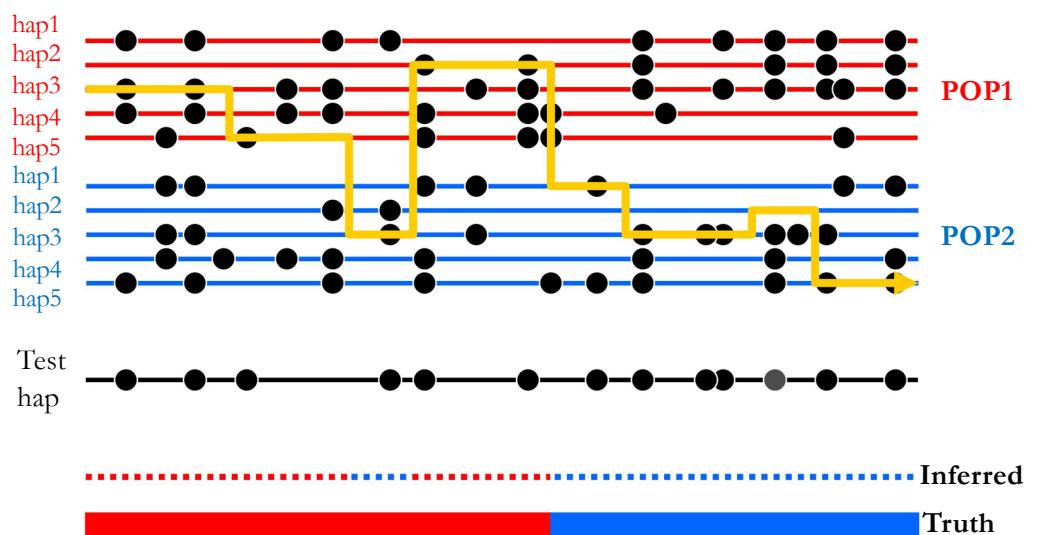
## Locally

- Generally supervised approach, i.e. leveraging reference panel of haplotypes assumed to be representative of the ancestral populations.
- *Generative* HMM type of approaches
- *Discriminative* conditional random field type of approaches

# Inferring local ancestry via HMM

HAPMIX (Price et al. PLoS Genet. 2009)

- Hidden states are local ancestry AND source haplotype from POP1 or POP2.
- So models both transitions between local ancestry states (Patterson et al. AJHG 2004) and between haplotypes from ancestral reference populations (Li & Stephens, Genetics 2003).
- Given initial, transition, and emission probabilities, use the forward-backward algorithm to infer  $P(\text{states} | \text{data})$



Advantage: really used all information available in GWAS array data, especially by using LD information.

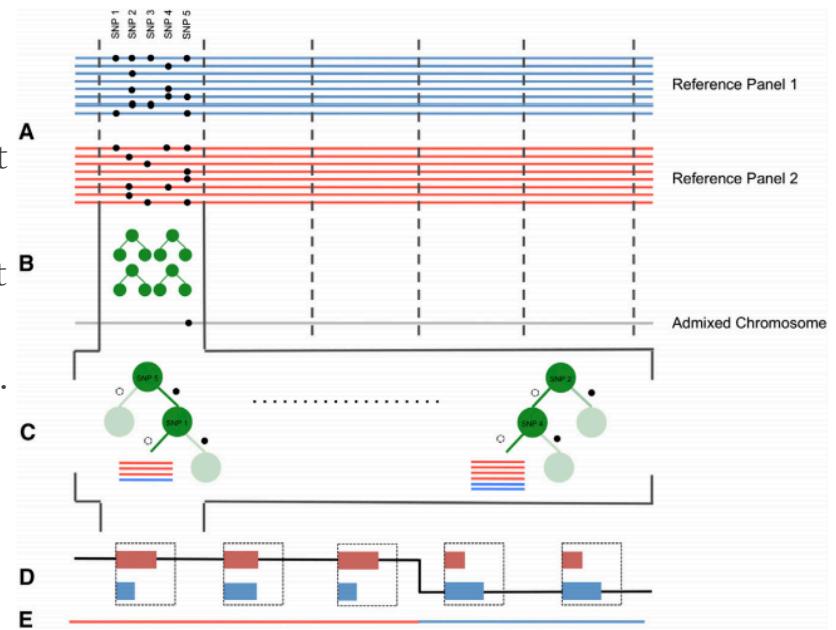
Disadvantage: increased complexity and computationally intensive. Limited to 2-way admixed populations.

# Inferring local ancestry via CRF

RFMix (Maples et al. AJHG 2013)

- "Supervised machine learning"
- Break genome into windows. In each window a random forest is trained to distinguish ancestry in the reference panels.
- Consider the test chromosome, each tree in the random forest generates a fractional vote for each ancestry by following the path through the tree corresponding to the admixed sequence.
- Votes are summed to produce posterior ancestry probabilities and most likely sequence of ancestry.

Much faster, can handle multi-way admixture ( $> 3$ )



# What are the uses for local ancestry?

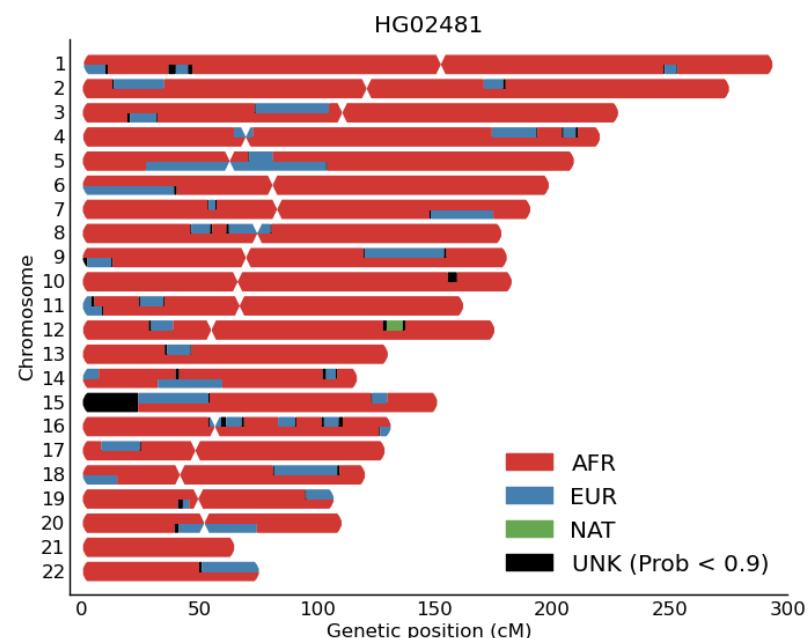
Understanding the demographic history of admixed populations

- Timing of admixture, number of pulses, sex-biased?

Detect adaptive introgression

Admixture mapping to detect disease alleles that might be particularly prevalent in one of the ancestral population.

- Particularly since the disease allele does not need to be genotyped – captured by the local ancestry.



[https://github.com/armartin/ancestry\\_pipeline](https://github.com/armartin/ancestry_pipeline)

# Population Genetic forces that shaped genetic variation

---

Demographic history

- Population structure
- Bottleneck
- Admixture

Natural Selection

# Natural Selection

---

**Selection** is a powerful force of evolution, and comes in many forms. Ultimately though, it acts on an individual's trait or traits to select the fit and remove the unfit.

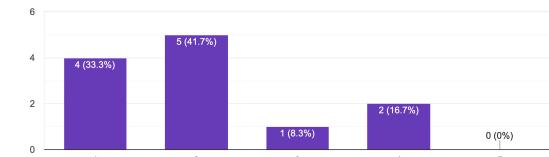
- What is “fitness”? Survival? Reproductive success?

When the selected trait is heritable in a population, natural selection would impact the **frequencies of alleles** underlying the selected trait.

Types of Selection:

- Positive selection or adaptation
- Negative or purifying selection
- ... and others, depending on whether one categorize by effect exerted on phenotype or genetic diversity.

13. How familiar are you with directional selection?  
12 responses



# A quick word on purifying selection

---

Most *de novo* mutations are deleterious rather than beneficial.

- Estimated **distribution of fitness effect (DFE)** on new mutations in the coding region of the human genome (assuming a European population demography) suggests ~30-33% of mutations are neutral or nearly neutral (~67-70% are deleterious to some extent). (Kim et al. Genetics 2017)

**Purifying selection** removing deleterious mutations help maintain the long-term stability of optimized biological structure. Purifying selection ensures that frequencies of deleterious alleles are low in the population.

- This contributed to the relationship between MAF and effect size in GWAS.

Purifying selection against a deleterious allele will also remove neutral variation linked to that allele, in what is known as **background selection**.

# Adaptation vs. purifying selection

---

**Positive selection** is usually thought of in the context of recent and local selective pressure that would work to increase the frequency of beneficial alleles. *i.e. adaptation.*

I tend to associate purifying selection with biology, which is shared among all human populations\*. Whereas adaptation is more a function of human population, where selective forces differ among populations (due to different environments).

A particular trait could be simultaneously subjected to purifying selection and positive selection (in a particular population around the world).

\*BUT, the efficiency of purifying selection could differ between populations, since population size influences the relative impact between drift and selection.

# Known examples of human adaptation?

Dairy consumptions

Arctic environment

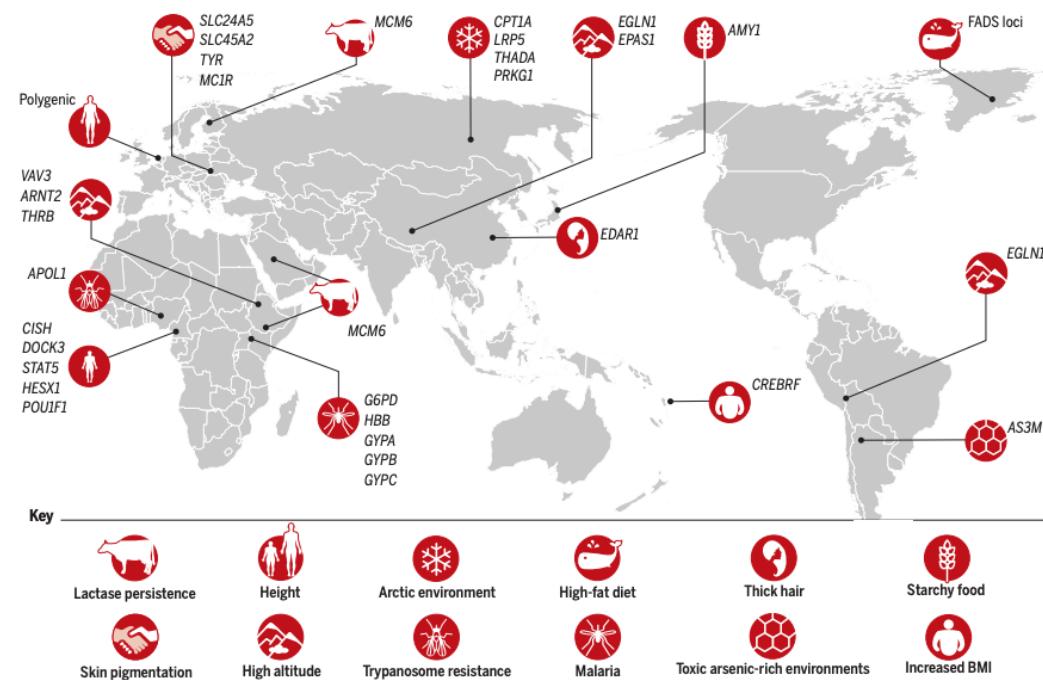
Endemic pathogens

High altitude

Toxic environment

UV exposure

...



Fan et al. Science 2016

# What was adaptative in the past could become maladaptive today and contribute to diseases

GDF5 variant affects bone growth, positively selected in cold climate, but increased arthritis risk.

APOL1 variant protects against African sleeping disease, positively selected in Africa, but increase kidney disease risk.

CREBRF variant lowers energy use and increase adipose storage, positively selected in Samoans, but increase obesity risk.

...

**ARTICLES**  
**nature genetics**

Ancient selection for derived alleles at a *GDF5* enhancer influencing human growth and osteoarthritis risk

Terence D Capellini<sup>1,2,7</sup>, Hao Chen<sup>2,6,7</sup>, Jiaxue Cao<sup>1,6,7</sup>, Andrew C Doxey<sup>3</sup>, Ata M Kiapour<sup>4</sup>, Michael Schoor<sup>2,6</sup> & David M Kingsley<sup>2,5</sup>

## Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans

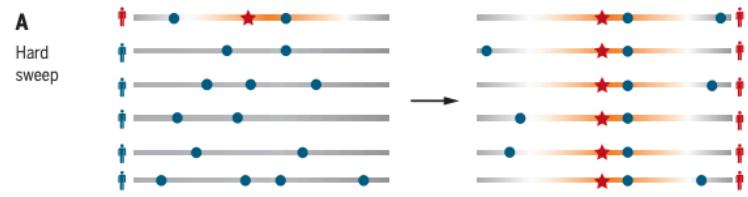
Giulio Genovese<sup>3,2\*</sup>, David J. Friedman<sup>1,3\*</sup>, Michael D. Ross<sup>4</sup>, Laurence Lecordier<sup>5</sup>, Pierrick Uzureau<sup>5</sup>, Barry I. Freedman<sup>6</sup>, Donald W. Bowden<sup>7,8</sup>, Carl D. Langefeld<sup>8,9</sup>, Taras K. Oleksyk<sup>10</sup>, Andrei L. Uscinski Knob<sup>4</sup>, Andrea J. Bernhardy<sup>1</sup>, Pamela J. Hicks<sup>7,8</sup>, George W. Nelson<sup>11</sup>, Benoit Vanhollebeke<sup>5</sup>, Cheryl A. Winkler<sup>12</sup>, Jeffrey B. Kopp<sup>11</sup>, Etienne Pays<sup>5,†</sup> & Martin R. Pollak<sup>1,13†</sup>

**LETTERS**  
**nature genetics**

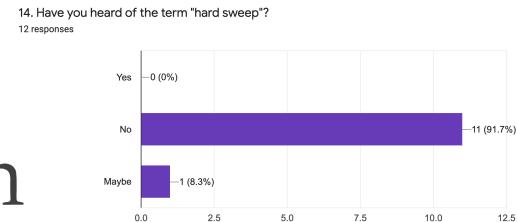
A thrifty variant in *CREBRF* strongly influences body mass index in Samoans

Ryan L Minster<sup>1,13</sup>, Nicola L Hawley<sup>2,13</sup>, Chi-Ting Su<sup>1,12,13</sup>, Guangyun Sun<sup>3,13</sup>, Erin E Kershaw<sup>4</sup>, Hong Cheng<sup>3</sup>, Olive D Buhule<sup>5,12</sup>, Jerome Lin<sup>1</sup>, Maugututia's Seifuva Reupena<sup>6</sup>, Sataupatea Viali<sup>7</sup>, John Tuiteme<sup>8</sup>, Take Naseri<sup>9</sup>, Zsolt Urban<sup>1,14</sup>, Ranjan Deka<sup>3,14</sup>, Daniel E Weeks<sup>1,5,14</sup> & Stephen T McGarvey<sup>10,11,14</sup>

# Genomic signature of adaptation



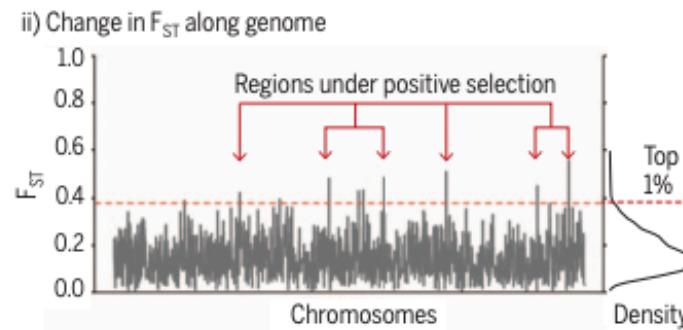
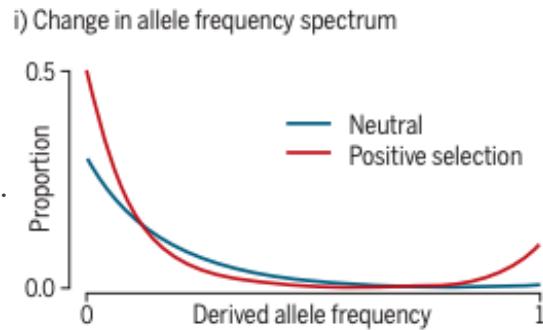
**Hard sweep:** a new advantageous *de novo* mutation quickly rise in frequency and fix in the population. Neutral variations linked to the advantageous allele will “**hitch-hike**” to higher frequency as well.



Fan et al. Science 2016  
Pritchard et al. Curr Bio 2010

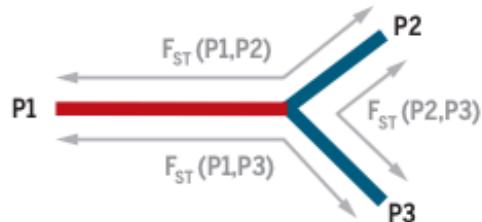
# Methods to detect recent positive selection

e.g. CLR (Nielsen et al. Genome Res. 2005)

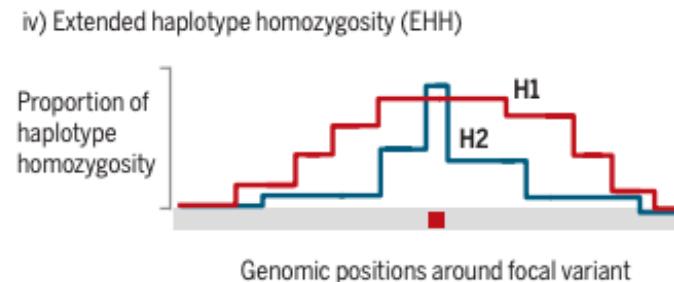


\*  $F_{ST}$ , between two pops, is equal to the proportion of genotypic variance in each pop attributable to population differences.

e.g. PBS (Yi et al. Science 2010)



Fan et al. Science 2016



e.g. xpEHH (Sabeti et al. Nature 2007), iHS (Voight et al. PLoS Biol 2006), nSL (Ferrer-Admetlla et al. MBE 2014) ...

# And many more...

---

Additional haplotype/genealogy-based tests:

SDS (Field et al. Science 2016)

ASMC (Palamara et al. Nat. Genet. 2018)

PRS trajectory (Edge and Coop, Genetics 2019)

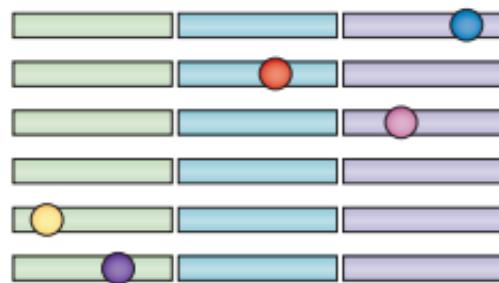
Relate (Speidel et al. Nat. Genet. 2019)

...

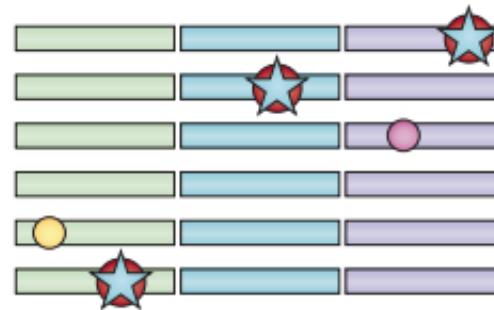
# Polygenic adaptation

## c Selection on a complex trait

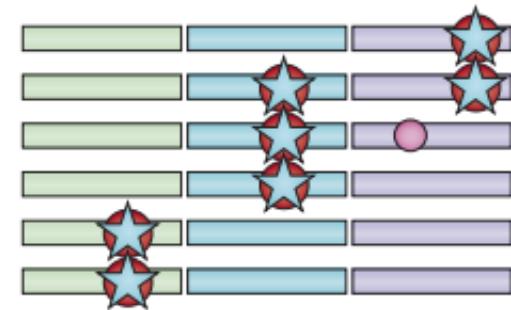
Neutral variation



A set of variants becomes adaptive in a new environment



Over time, the set of variants becomes more common

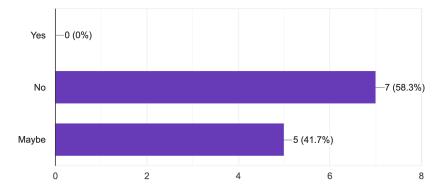


Pritchard & Di Rienzo, Nat Rev Genet 2010

Pritchard et al. Curr Bio 2010

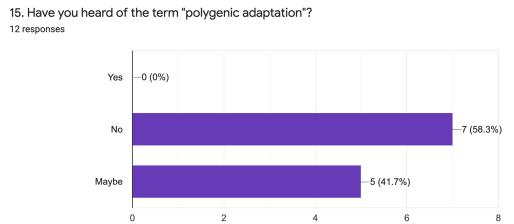
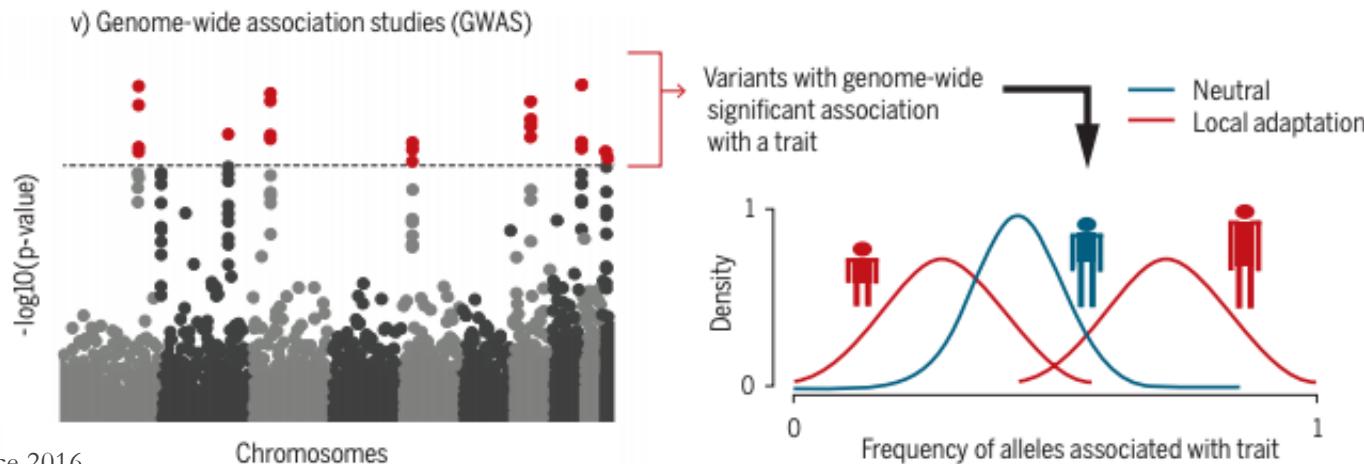
Scheinfeldt & Tishkoff, Nat Rev Genet 2013

15. Have you heard of the term "polygenic adaptation"?  
12 responses



# Polygenic adaptation

Haplotype signature VERY SUBTLE. But we can take a more phenotypic-driven approach and look for concerted directional shift in frequency of GWAS SNPs (e.g. trait-increasing alleles tend to increase in frequency in the taller-trait value population).



e.g. Turchin, Chiang et al. (Nat. Genet. 2012), Berg and Coop (PLoS Genet. 2014), Robinson et al. (Nat. Genet. 2015), Racimo et al. (Genetics 2018) ...

# Population Genetic forces that shaped genetic variation

---

## ~~Demographic history~~

- ~~Population structure~~
- ~~Bottleneck~~
- ~~Admixture~~

## ~~Natural Selection~~

**BONUS MATERIAL** analysis of polygenic adaptation confounded by uncorrected population structure

# Detection of polygenic adaptation relies on unbiased GWAS results

---

But GWAS could be confounded by a phenomenon known as **populations stratification** (i.e. population substructure that was uncorrected in analysis).

# Successful-Use-of-Selected-Hand- Instruments (SUSHI) gene



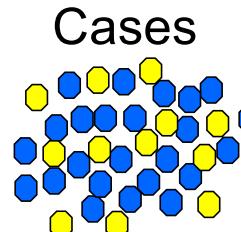
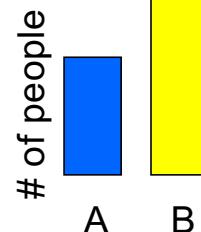
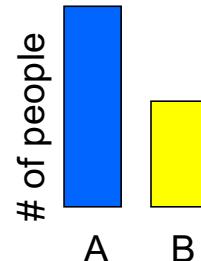
Study population is diverse,  
unaware of substructure



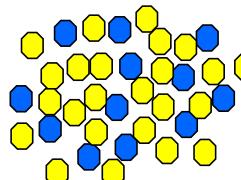
Trait is chopsticks use



## Prevalence



## Controls



$$\sim \alpha_0 + \alpha_1 * \text{genotype}$$

Lander & Schork, Science 1994  
Hamer & Sirota, Mol. Psych. 2000

# ~~Successful-Use-of-Selected-Hand-Instruments (SUSHI) gene~~



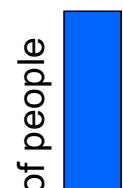
Study population is diverse,  
unaware of substructure



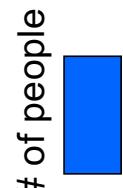
Trait is chopsticks use



## Prevalence



A



B

## Cases

Any variant with significantly different frequency between subpopulations A and B, due to their different population history and drift, will appear to be associated with the trait of interest. Vast majority will be spurious. type

Lander & Schork, Science 1994  
Hamer & Sirota, Mol. Psych. 2000

Adult height is differentiated across Europe

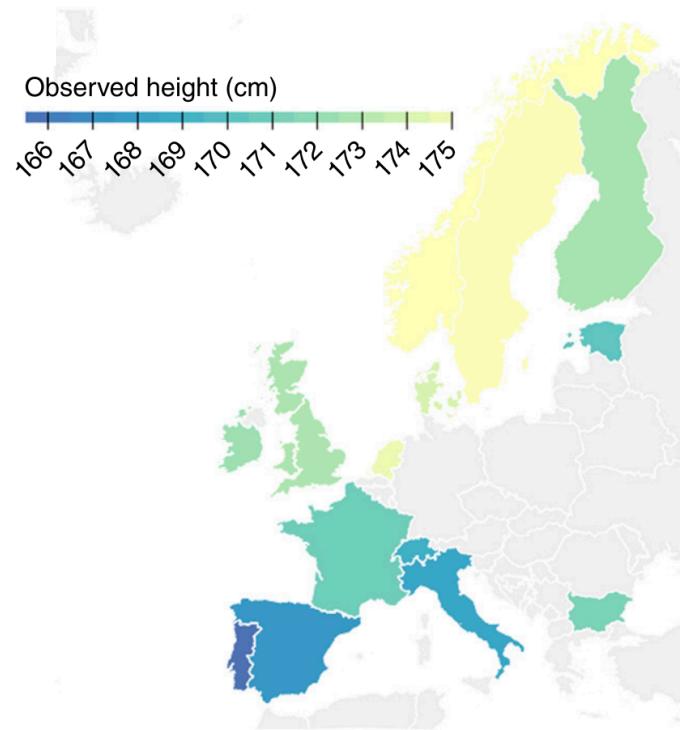
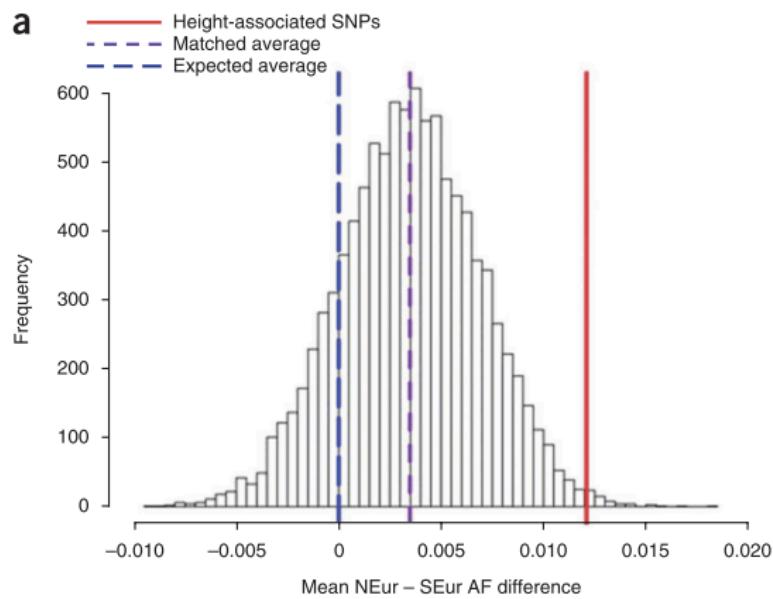


Figure from Robinson *et al.*, Nat. Genet. 2015



Is this difference across populations driven by natural selection?

# At height-associated loci, height-increasing allele are more common in Northern EUR



## Evidence of widespread selection on standing variation in Europe at height-associated SNPs

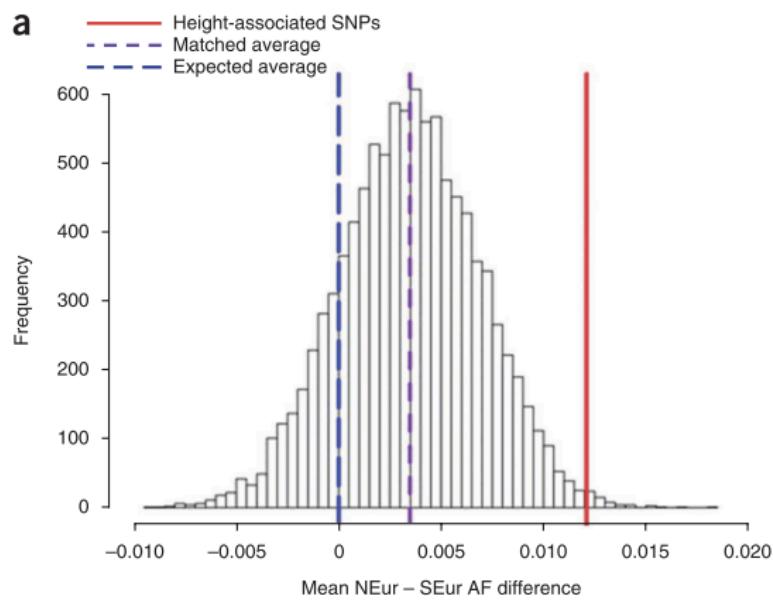
Michael C Turchin<sup>1-5,8</sup>, Charleston WK Chiang<sup>1-6,8</sup>, Cameron D Palmer<sup>1-5</sup>, Sriram Sankararaman<sup>5,6</sup>, David Reich<sup>5,6</sup>, Genetic Investigation of ANthropometric Traits (GIANT) Consortium<sup>7</sup> & Joel N Hirschhorn<sup>1-6</sup>

61% (85/139) of height-increasing alleles have higher frequency in NEur ( $P = 0.011$ )

Frequencies of height-increasing alleles are on avg 1.2% higher in NEur vs. SEur ( $P = 4.3 \times 10^{-4}$ )

This pattern is far more likely in model with selection than in model with genetic drift alone ( $P = 10^{-15}$ )

# At height-associated loci, height-increasing allele are more common in Northern EUR



61% (85/139) of height-increasing alleles have higher frequencies in Northern EUR than expected under uncorrected population structure. This pattern is far more likely in model with selection than in model with genetic drift alone ( $P = 10^{-15}$ )

More height loci or uncorrected population structure? ( $P = 0.011$ )

Northern EUR ( $P = 4.3 \times 10^{-4}$ )

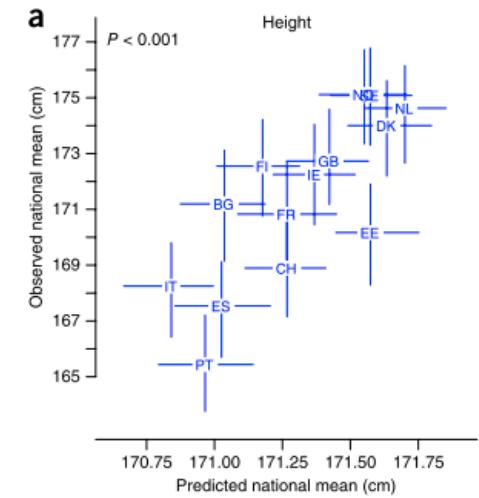
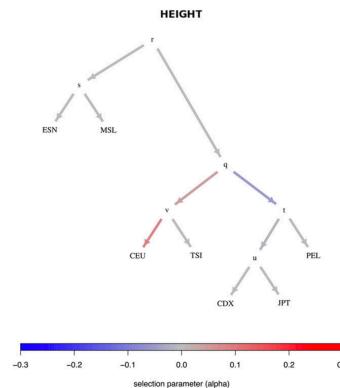
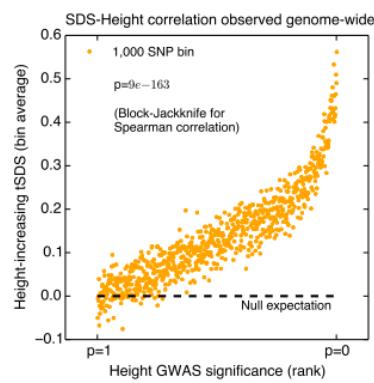
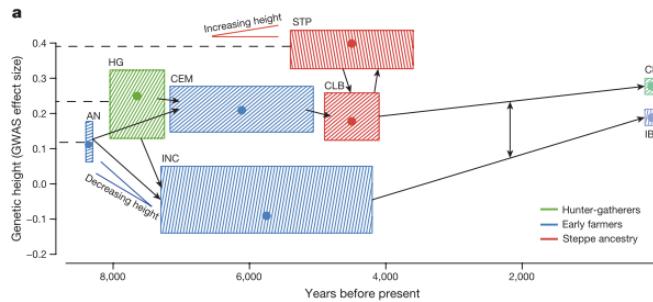
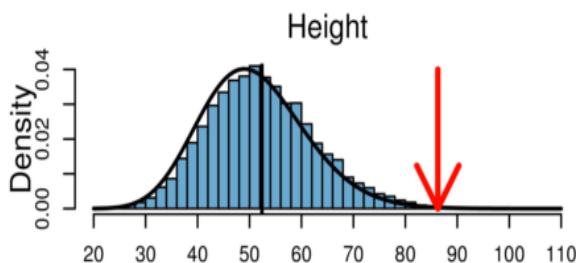
LETTERS

nature  
genetics

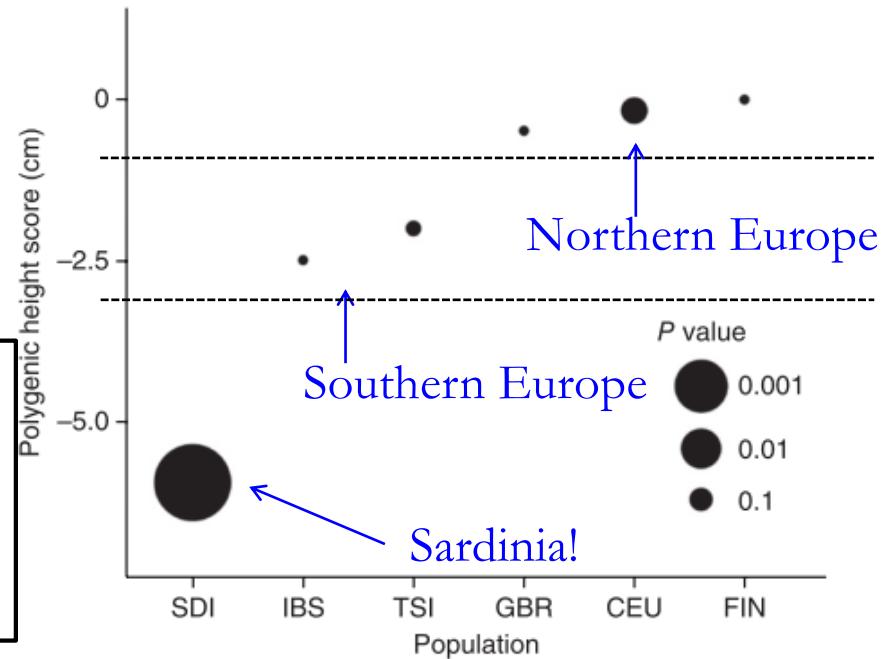
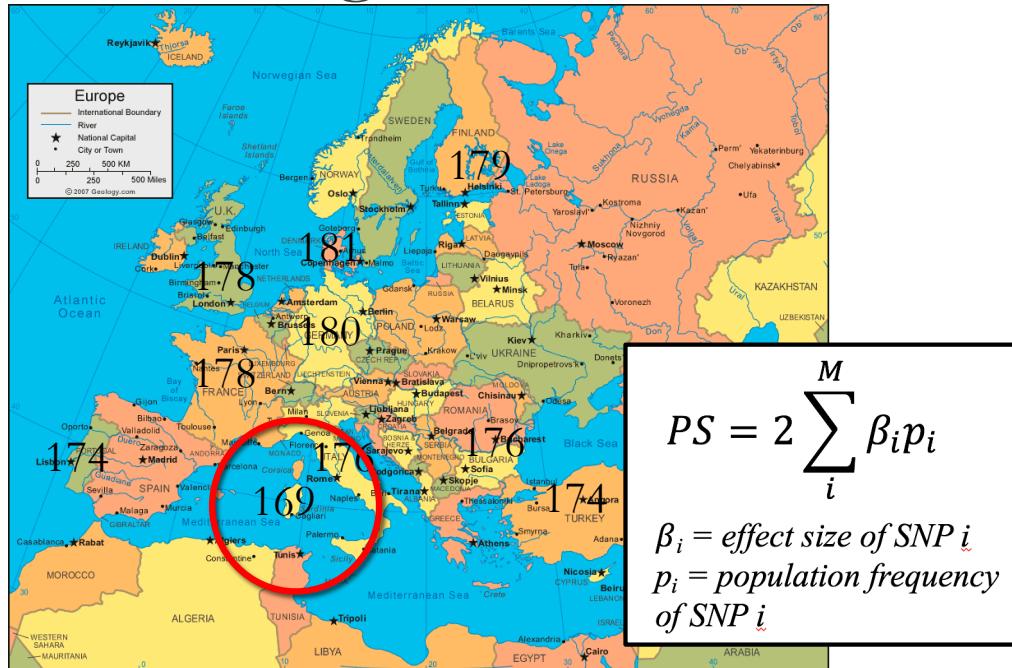
## Evidence of widespread selection on standing variation in Europe at height-associated SNPs

Michael C Turchin<sup>1-5,8</sup>, Charleston WK Chiang<sup>1-6,8</sup>, Cameron D Palmer<sup>1-5</sup>, Sriram Sankararaman<sup>5,6</sup>, David Reich<sup>5,6</sup>, Genetic Investigation of ANthropometric Traits (GIANT) Consortium<sup>7</sup> & Joel N Hirschhorn<sup>1-6</sup>

# Sparked a field of method development in detecting polygenic adaptation



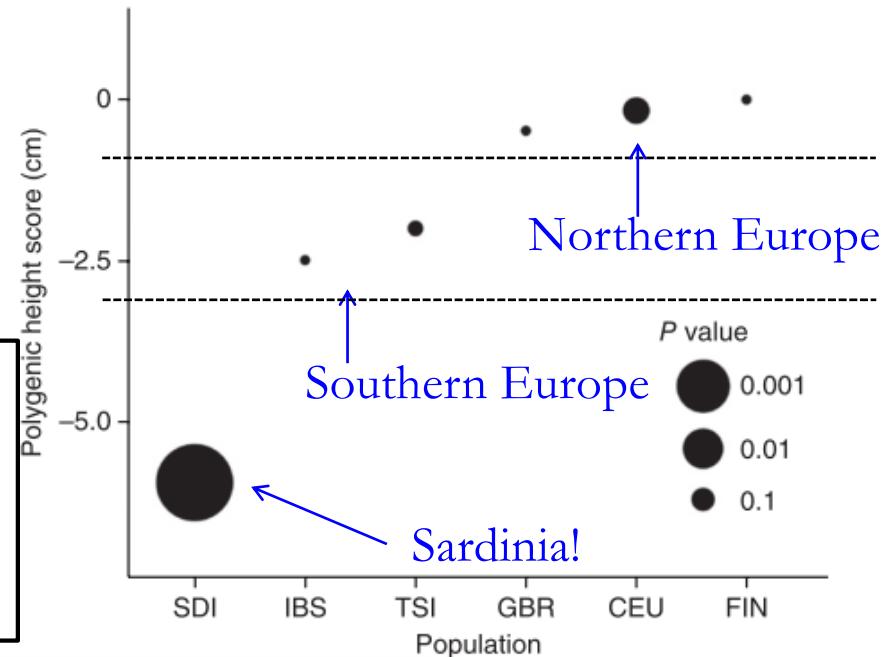
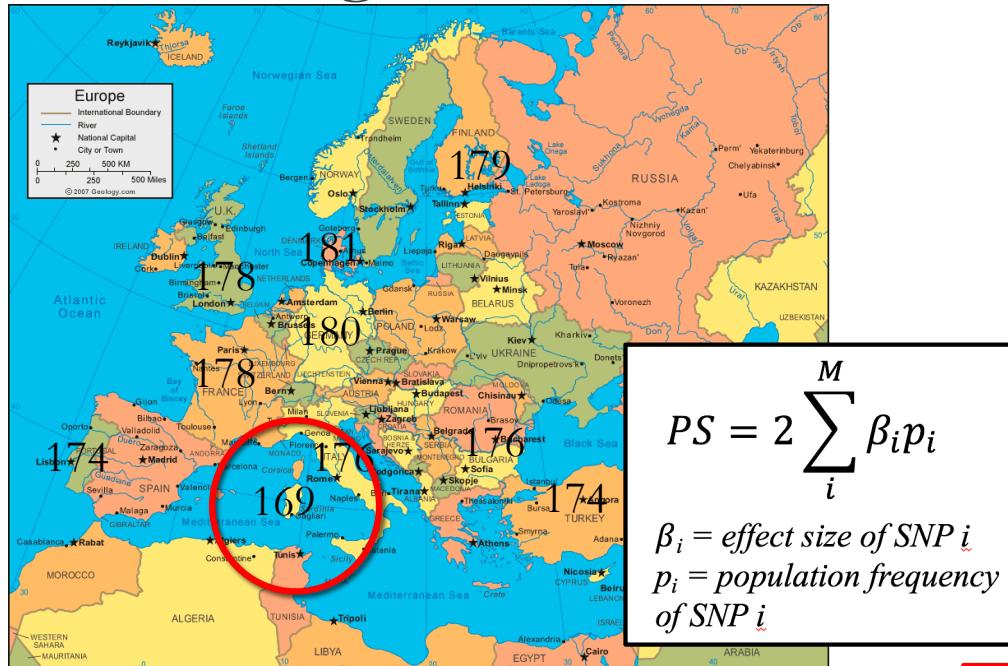
# Shorter height more favorable in Sardinia



Based on polygenic height scores, Sardinians are on average ~6 cm shorter than CEU. Conditional on the degree of genetic similarity between Sardinians and Europeans, this difference is significant ( $P = 1.6 \times 10^{-6}$ ). Height-decreasing alleles are on average 1.6% more prevalent in Sardinia than in reference Northern European.

Zoledziewska\*, Sidore\*, Chiang\*, Sanna\* et al. Nat. Genet. 2015

# Shorter height more favorable in Sardinia



Based on polygenic height scores, Sardinians are on average  $\sim 6$  cm shorter than CEU. Conditional on the degree of genetic similarity between Sardinians and Europeans, this difference is significant ( $P = 1.6 \times 10^{-6}$ ). Height-decreasing alleles are on average 1.6% more prevalent in Sardinia than in reference Northern European.

Zoledziewska\*, Sidore\*, Chiang\*, Sanna\* et al. Nat. Genet. 2015

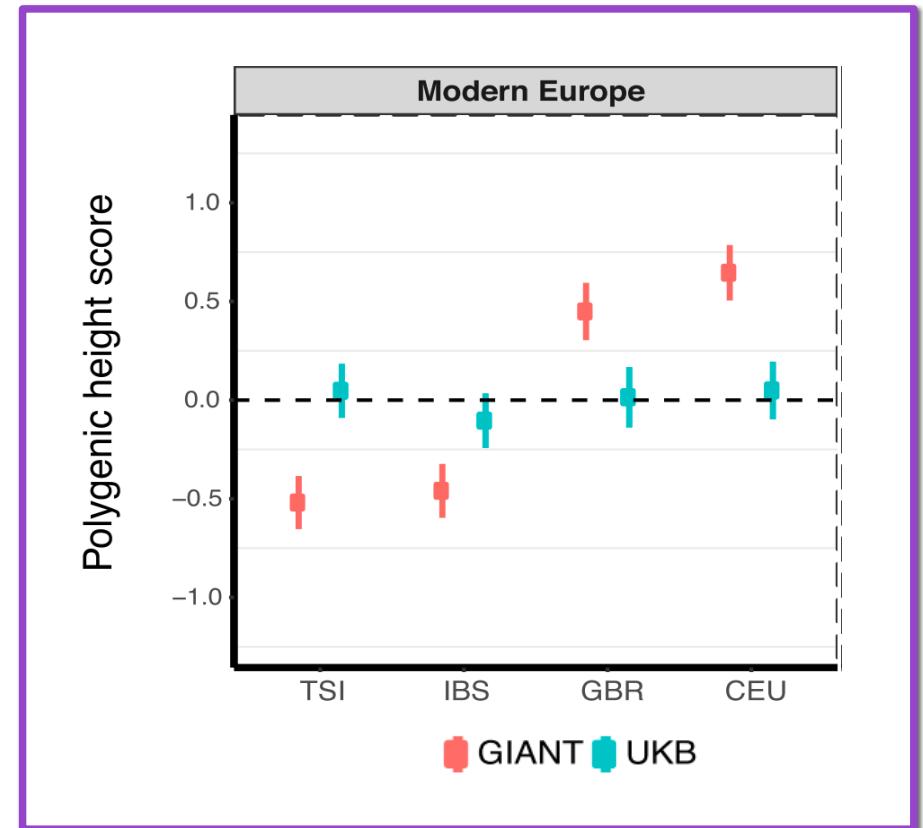


## Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies

Mashaal Sohail<sup>1,2,3†\$\*</sup>, Robert M Maier<sup>3,4,5‡\*</sup>, Andrea Ganna<sup>3,4,5,6,7</sup>,  
 Alex Bloemendaal<sup>3,4,5</sup>, Alicia R Martin<sup>3,4,5</sup>, Michael C Turchin<sup>8,9</sup>,  
 Charleston WK Chiang<sup>10</sup>, Joel Hirschhorn<sup>3,11,12</sup>, Mark J Daly<sup>3,4,5,7</sup>,  
 Nick Patterson<sup>3,13</sup>, Benjamin Neale<sup>3,4,5§\*</sup>, Iain Mathieson<sup>14‡\$\*</sup>, David Reich<sup>3,13,15‡\$\*</sup>,  
 Shamil R Sunyaev<sup>2,3,16‡\$</sup>

## Reduced signal for polygenic adaptation of height in UK Biobank

Jeremy J Berg<sup>1†\$\*</sup>, Arbel Harpak<sup>1,2†</sup>, Nasa Sinnott-Armstrong<sup>3†</sup>,  
 Anja Moltke Joergensen<sup>4</sup>, Hakhamaresh Mostafavi<sup>1</sup>, Yair Field<sup>3</sup>,  
 Evan August Boyle<sup>3</sup>, Xinjun Zhang<sup>5</sup>, Fernando Racimo<sup>4</sup>, Jonathan K Pritchard<sup>2,6\*</sup>,  
 Graham Coop<sup>7,8\*</sup>



# Is height difference across populations driven by natural selection?



BIOBANK JAPAN

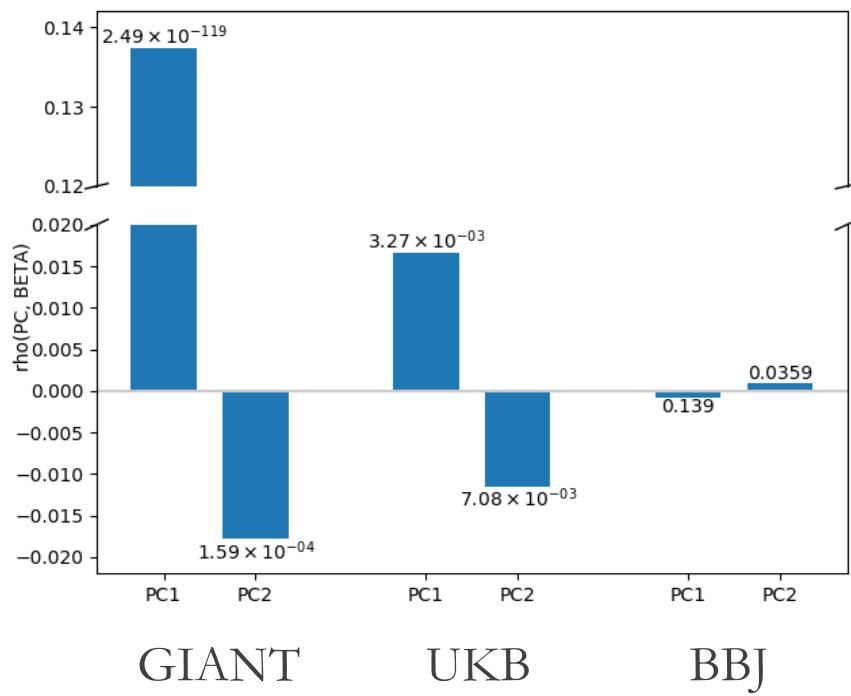
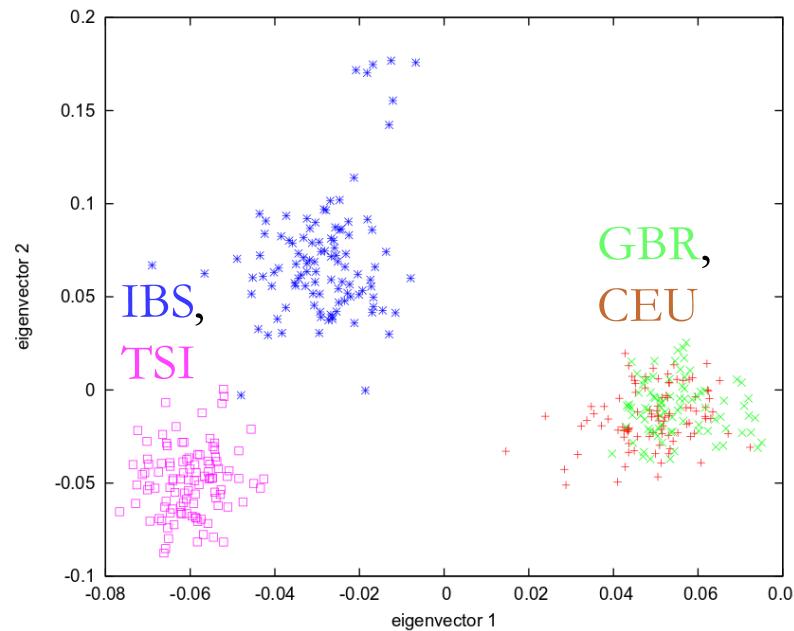
**BBJ**

N = 159K individuals,  
genome sequencing +  
imputation for 27.2M  
variants

(Akiyama et al. Nat. Comm. 2019)

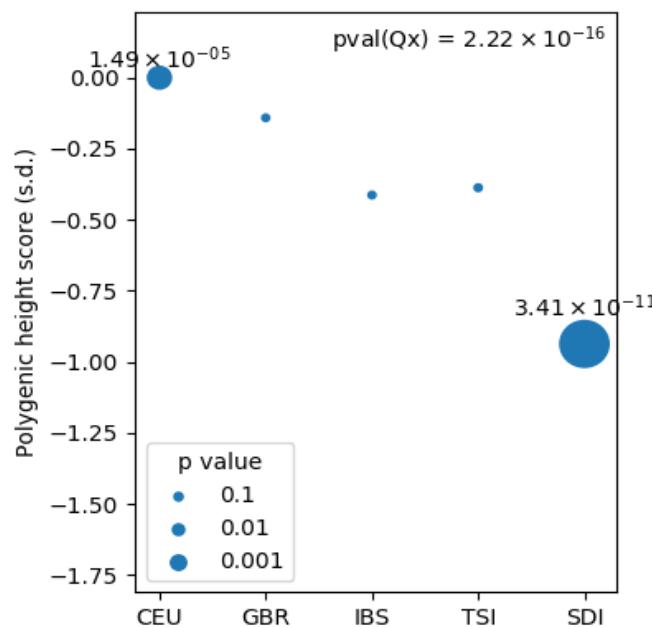
Minhui Chen, Ph.D.

Height loci ascertained from BBJ is (nearly) completely free of stratification in Europe



Chen et al, bioRxiv 2019

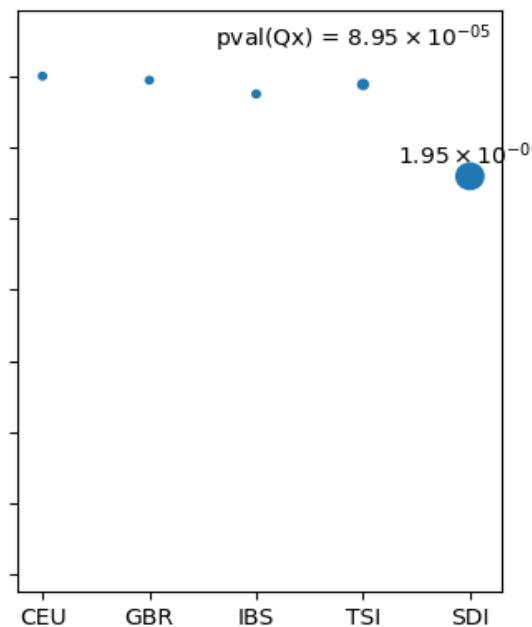
# Sardinians are still unexpectedly shorter than mainland Europeans, consistent with selection



Height-associated loci  
ascertained from:

Chen et al, bioRxiv 2019

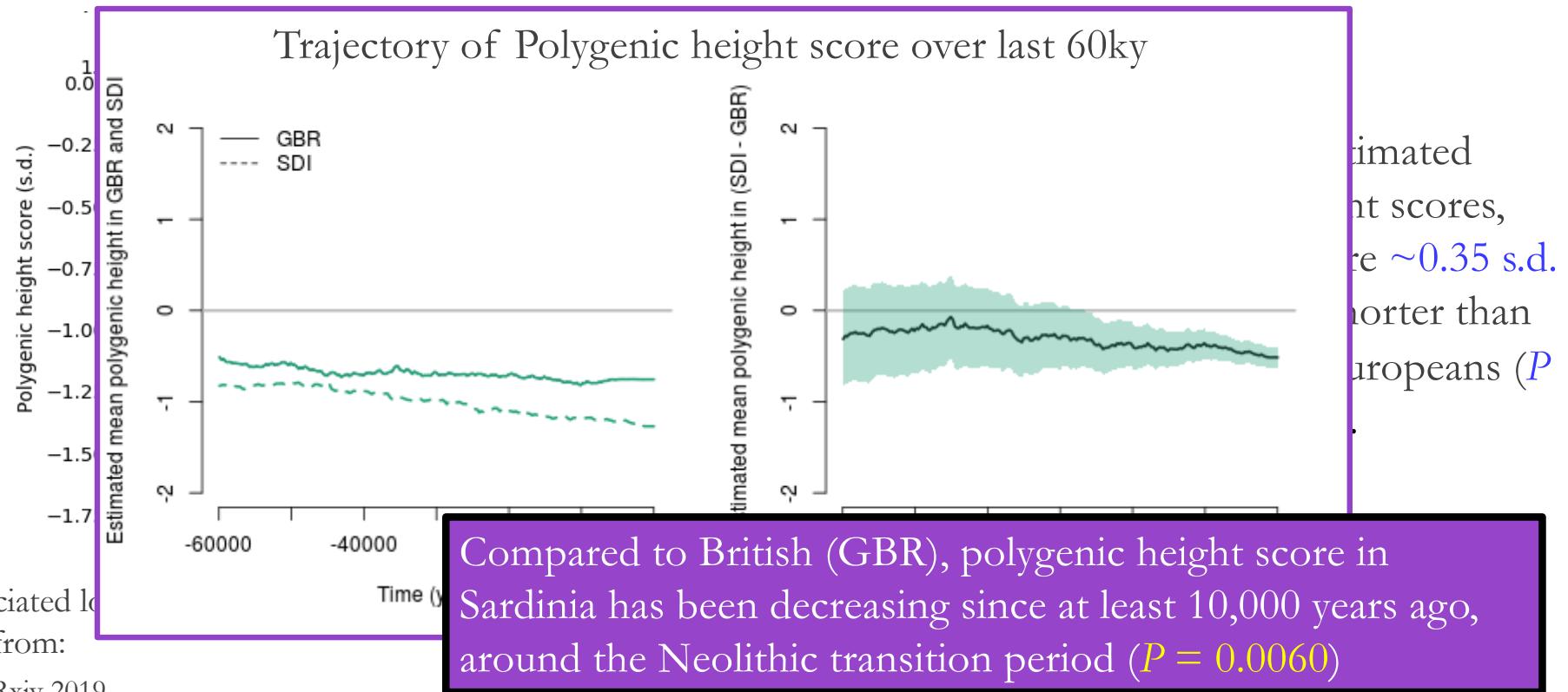
GIANT



BBJ

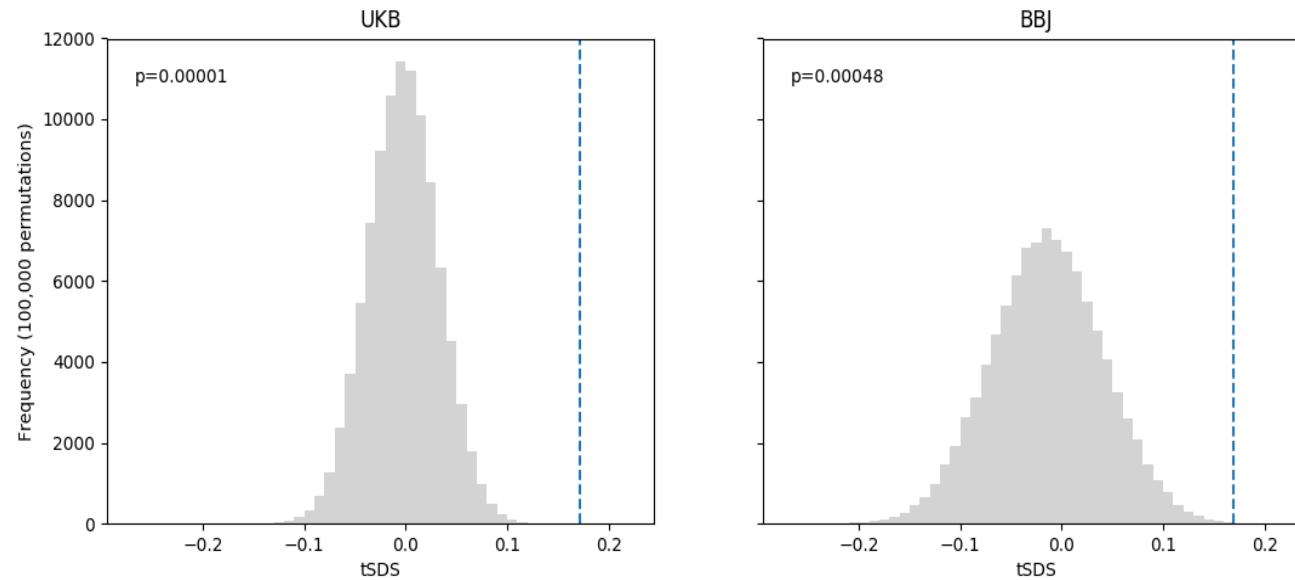
Based on estimated  
genetic height scores,  
Sardinians are  $\sim 0.35$  s.d.  
( $\sim 2.1$  cm) shorter than  
Northern Europeans ( $P$   
 $= 1.95 \times 10^{-6}$ ).

# Sardinians are still unexpectedly shorter than mainland Europeans, consistent with selection



In mainland Europe, the signal is much weaker...

... until we examined haplotype-based statistic, tSDS,  
from  $\sim 3000$  WGS British individuals



Using height-loci ascertained from BBJ, height-increasing alleles are significantly more likely to sit on a younger/longer haplotype ( $P = 4.8 \times 10^{-4}$ ), consistent with recent selection towards taller height in British.

Chen et al, bioRxiv 2019

# Conclusions

---

Carefully conducted GWAS is essential for understanding biology and evolution. Even very subtle biases could confound studies of natural selection like the height story.

We ascertained trait-associated loci from East Asian population to alleviate concerns of uncorrected stratification in testing adaptation.

Shorter height in Sardinia appears driven by selection, perhaps over the last 10,000 years.

Taller height could be due to selection in recent Northern Europeans, but signal is weaker than previously thought.

# Contact and Resources

---

Any questions or interests in these research themes?

- [charleston.chiang@med.usc.edu](mailto:charleston.chiang@med.usc.edu)
- <http://chianglab.usc.edu>

Other online resources and material (in which this lecture is heavily influenced by):

- <https://www.hsph.harvard.edu/alkes-price/epi511/>
- <https://github.com/cooplab/popgen-notes>
- [https://github.com/NovembreLab/HGDP\\_PopStruct\\_Exercise](https://github.com/NovembreLab/HGDP_PopStruct_Exercise)
- [https://github.com/NovembreLab/1000genomes\\_Selection\\_Exercise](https://github.com/NovembreLab/1000genomes_Selection_Exercise)