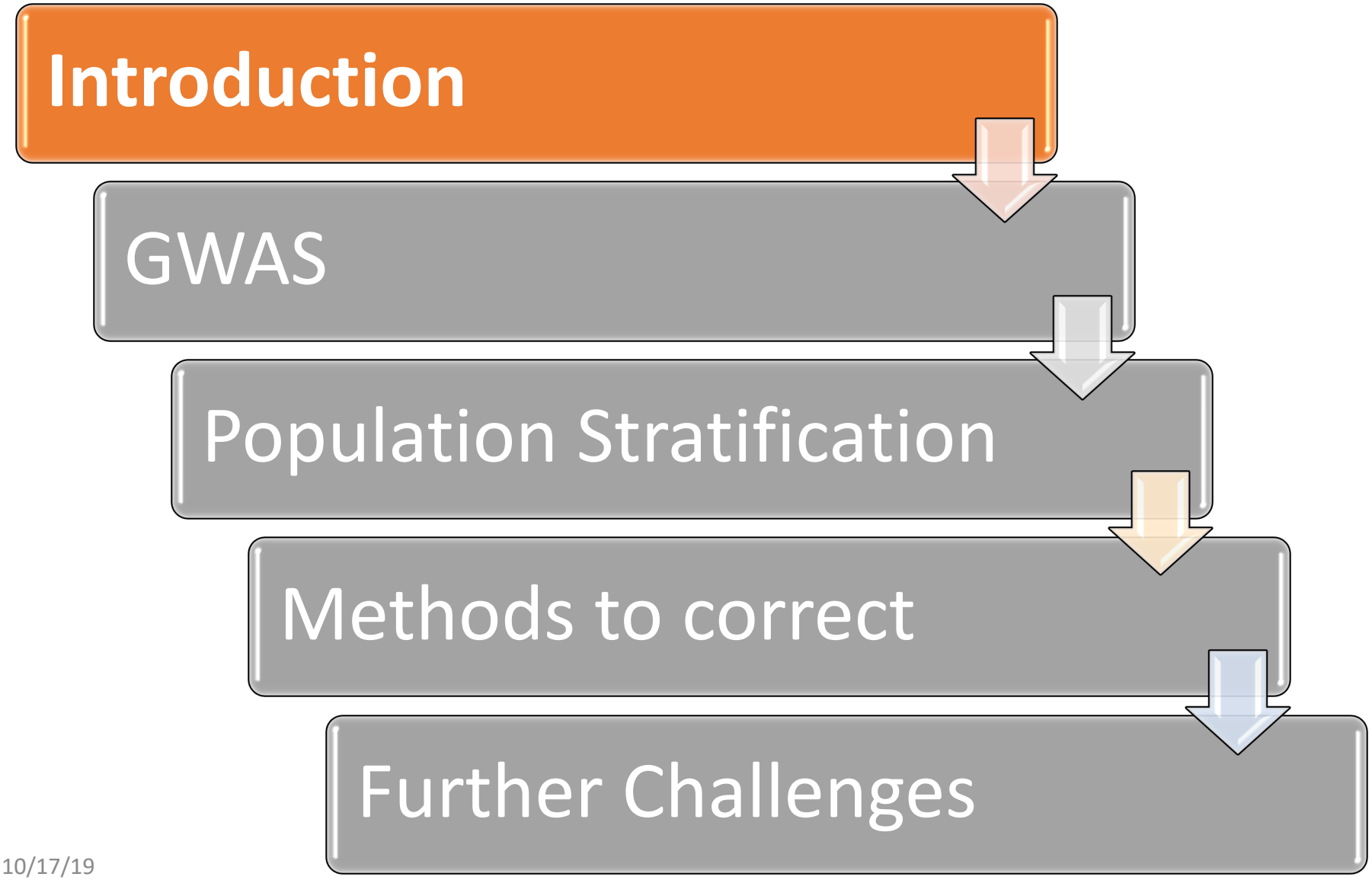
Population Stratification in Genome-Wide Association Studies

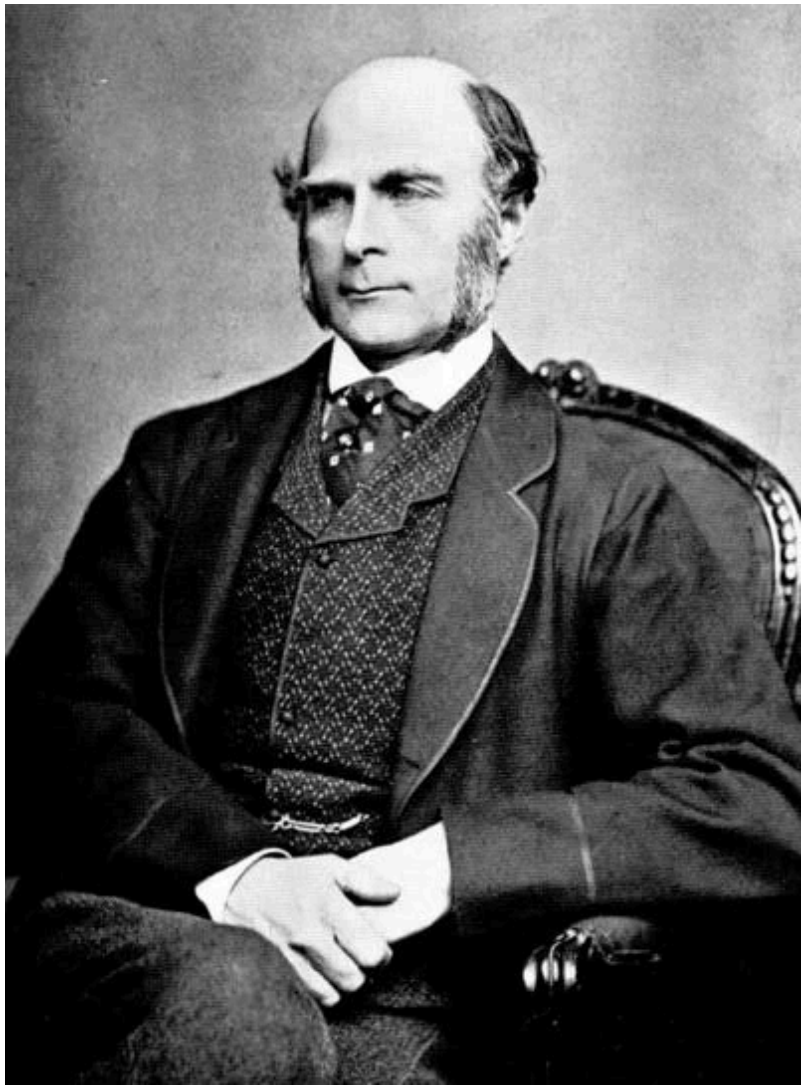
Jinyuan Qi

SML510: CSML Graduate Research Seminar

Oct 17th, 2019

Outline


Introduction
GWAS
Population Stratification
Methods to correct
Further Challenges



"Everything from criminality to love of poetry was thought to be in the hereditary nature of humans."

– from Francis Galton's <*Hereditary Genius*> 1869.

Horrors of Eugenics

- Forced sterilization of the "unfit"
- American racism
- Nazism

Motivation, Doubts & Criticism

NEWS GENETICS

There's no evidence that a single 'gay gene' exists

Instead, a combo of small genetic factors and environmental influences affect partner choice

How scientists are learning to predict your future with your genes

But what are the limits?

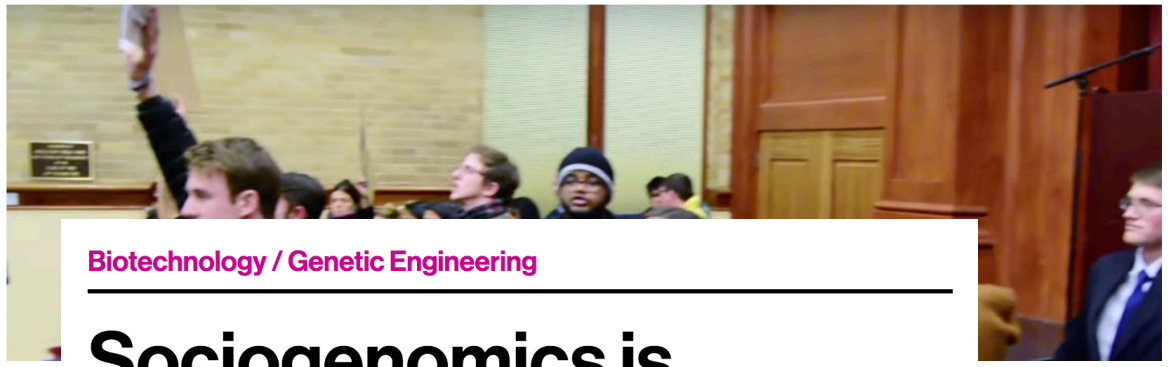
By Brian Resnick | @B_resnick | brian@vox.com | Updated Aug 25, 2018, 9:35am EDT

Graphics and illustrations by Javier Zarracina

Published on October 15, 2018

Is Sociogenomics Racist?

written by **Toby Young**



Biotechnology / Genetic Engineering

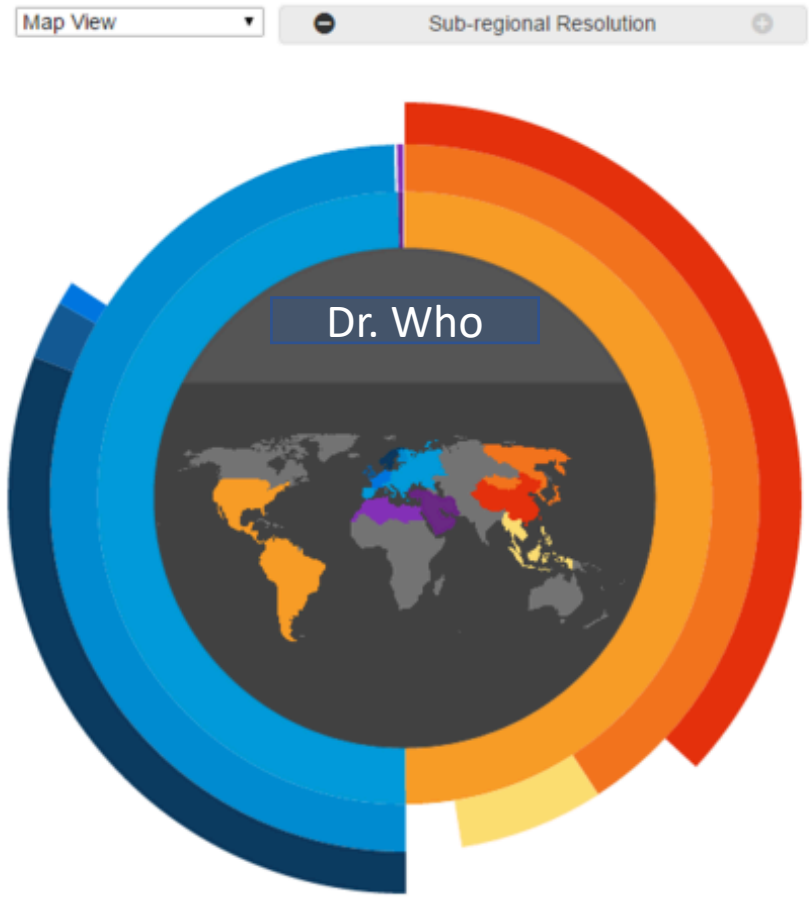
Sociogenomics is opening a new door to eugenics

New ways of using your genetic data could bolster scientific racism and encourage discrimination.

by **Nathaniel Comfort** Oct 23, 2018

Ref: Ganna, A., Verweij, K. J., Nivard, M. G., Maier, R., Wedow, R., Busch, A. S., ... & Lundström, S. (2019). Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science*, *365*(6456), eaat7693.

Personalized Genotyping – ancestry, health & traits



Health Predisposition

Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. This analysis includes DNA you received from all of your recent ancestors, on both sides of your family. The results reflect where your ancestors lived before the widespread migrations of the past few hundred years.

49.9%	East Asian & Native American	
	East Asian	
36.9%	Chinese	
3.9%	Broadly East Asian	
6.6%	Southeast Asian	
2.5%	Broadly East Asian & Native American	
49.7%	European	
	Northwestern European	
30.9%	Scandinavian	
2.4%	British & Irish	
1.0%	French & German	
15.4%	Broadly Northwestern European	
0.1%	Broadly European	
0.3%	Middle Eastern & North African	
0.3%	North African	
0.1%	Unassigned	

Dr. Who

show all populations

Carrier Status

Learn whether you have specific genetic variants that may not affect your health, but could affect your children's health.

REPORTS INCLUDE

Bloom Syndrome	Variant not detected
Cystic Fibrosis	Variant not detected

View 44 reports

Traits

Explore the genetics behind your appearance and senses.

View 34 reports

A primer in Genetics

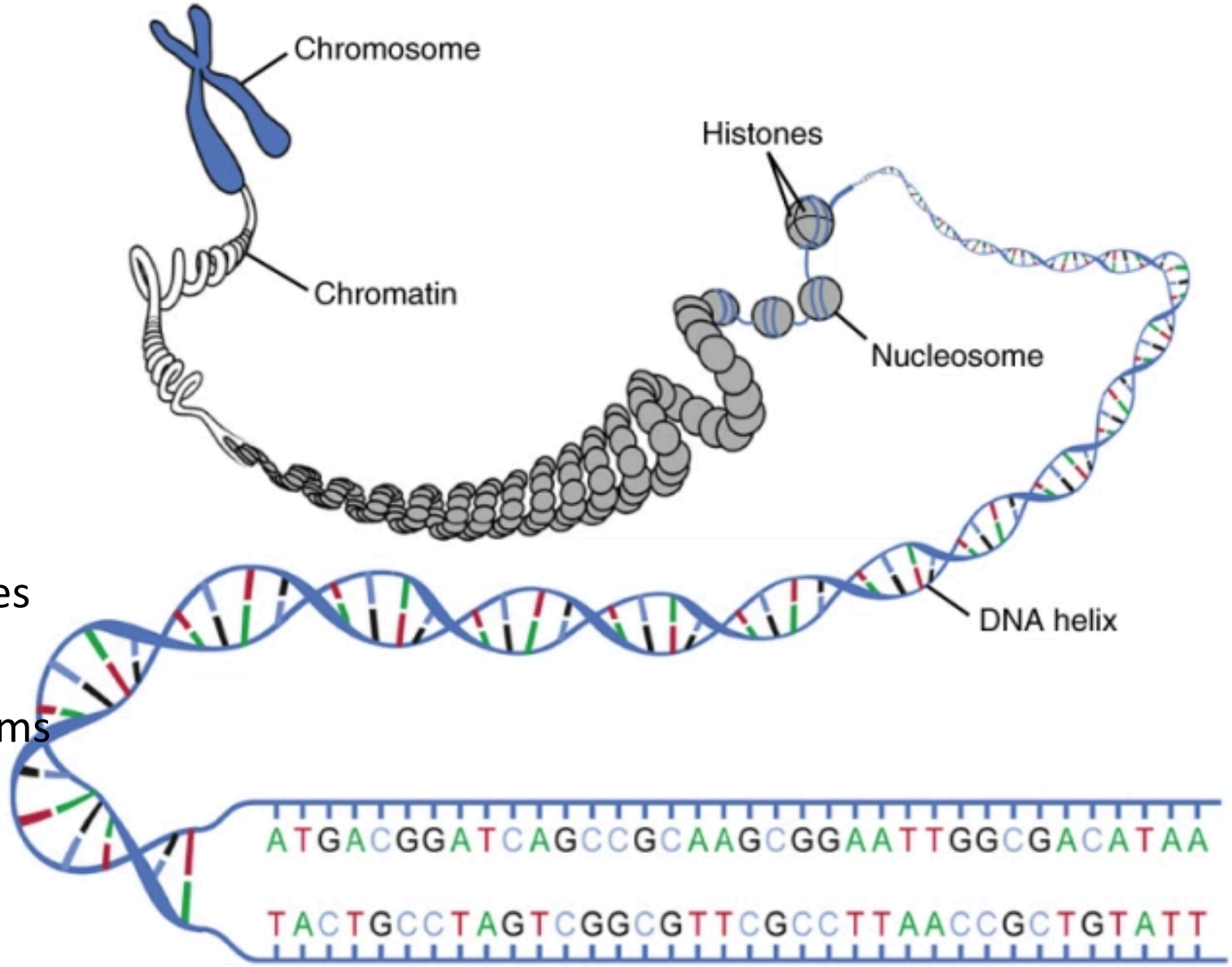
Humans

- 23 chromosome pairs
- ~20,000-25,000 genes
- ~3,200 Mb (megabase pairs)

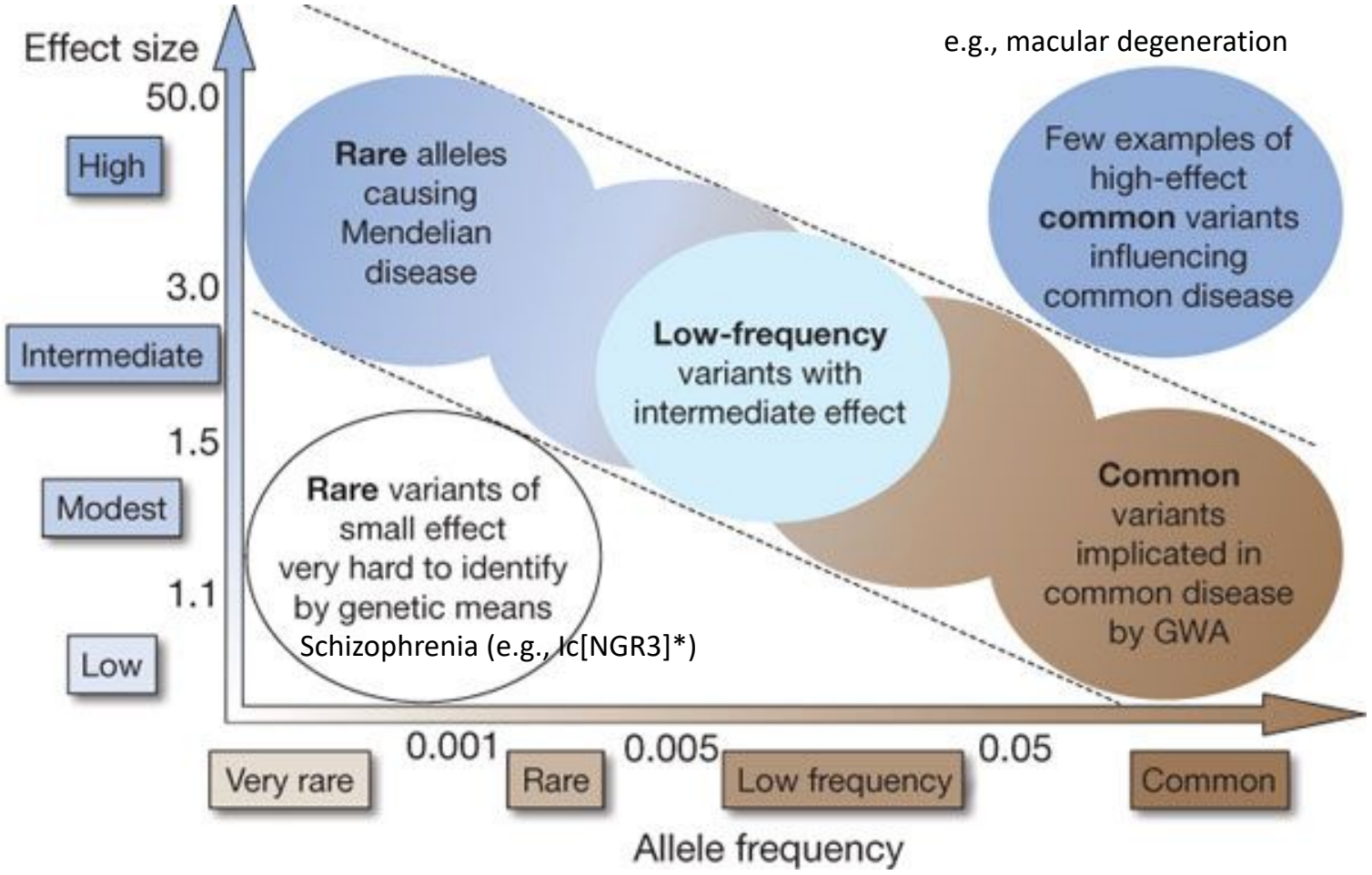
Genetic polymorphism

- **Alleles** - variations of a locus that codes for protein
- **SNPs** – single-nucleotide polymorphisms https://www.snpedia.com/
- **MAF** – minor allele frequency.

(common >0.05; rare <0.01)



The spectrum of genetic contribution to phenotypes



e.g., macular degeneration

- **Mendelian traits** - (e.g., sickle cell anemia, Huntington's disease)
- **Rare to low frequencies & moderate effects** (e.g., Crohn's disease)
- **Common variants will almost always have small effects** (e.g., diabetes, educational attainment, age at first birth.)

Monogenic (rare)

↓

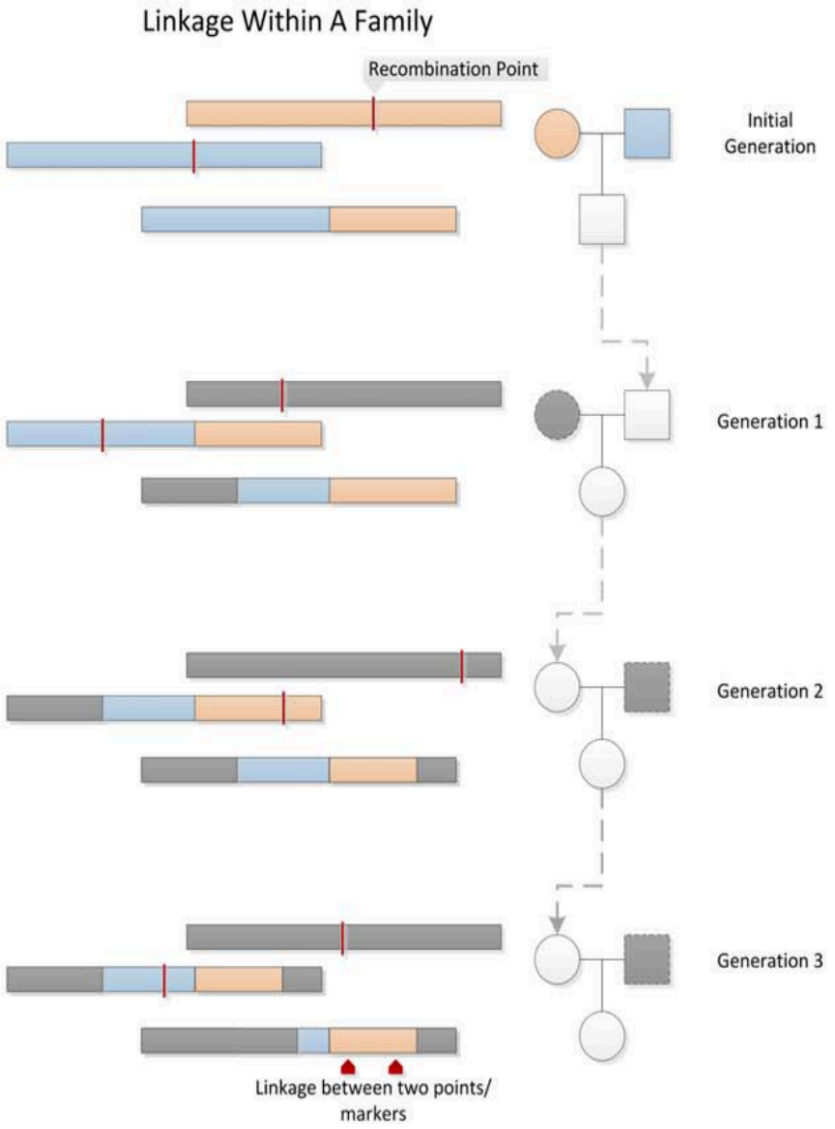
Polygenic

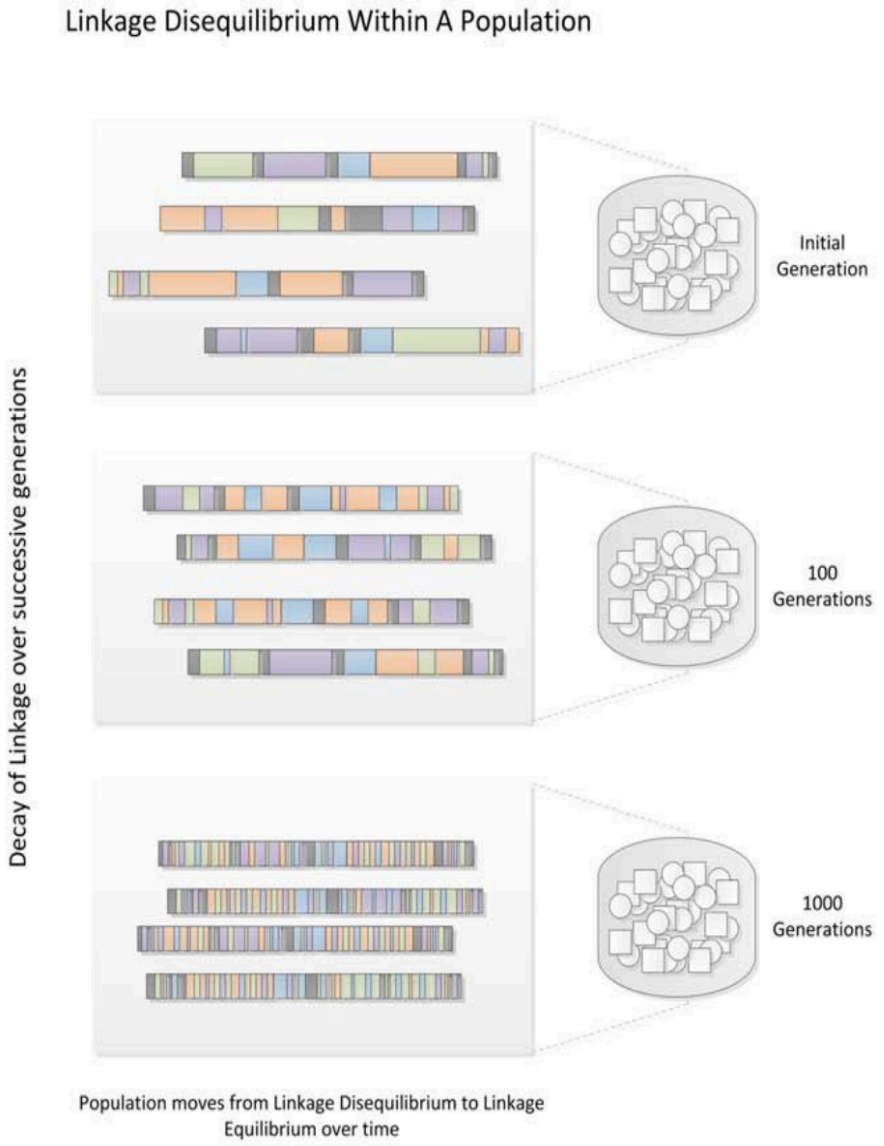
↓

Omnigenic

Ref: Manolio, Teri A., et al. "Finding the missing heritability of complex diseases." *Nature* 461.7265 (2009): 747.

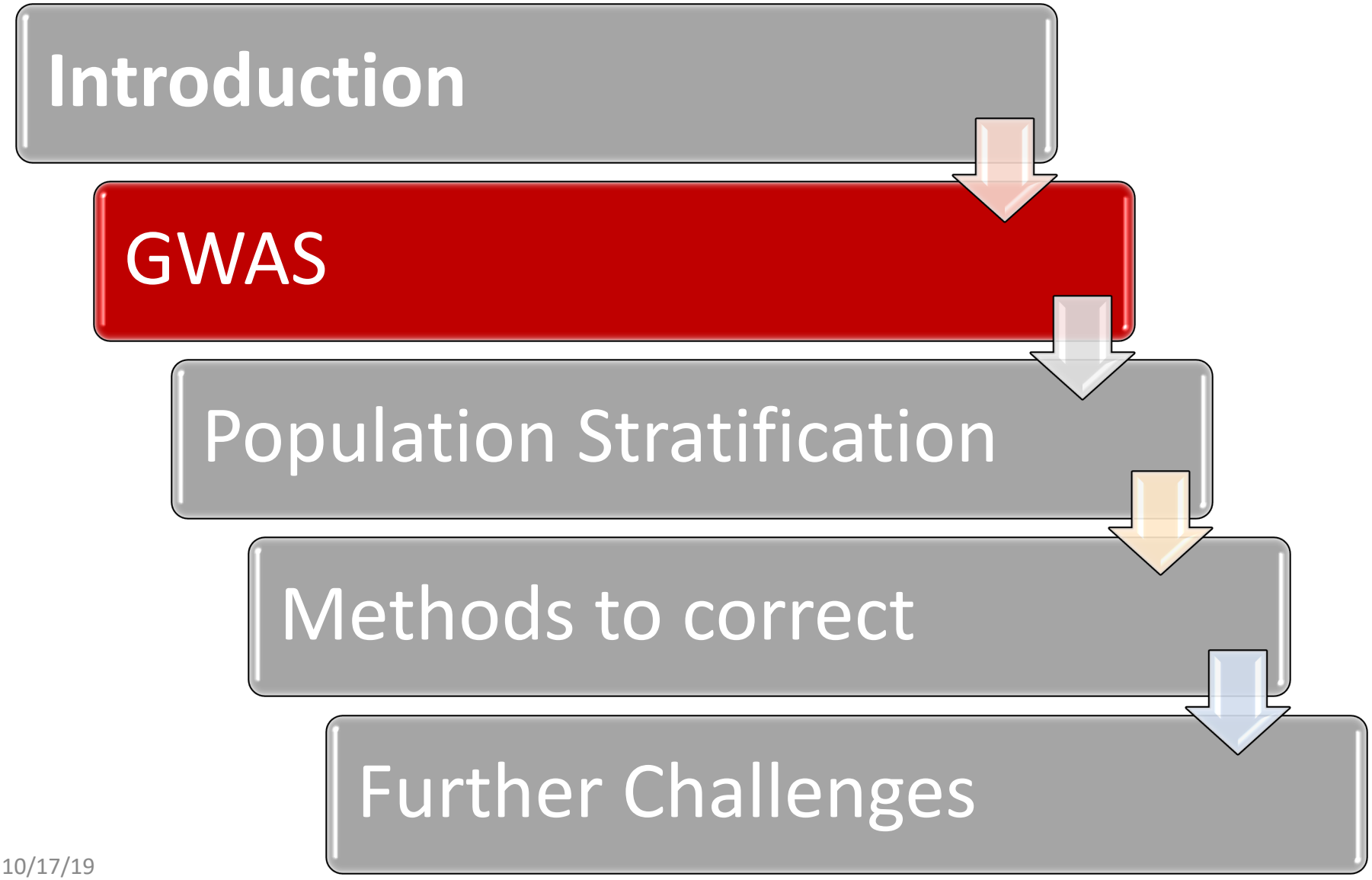
Linkage and Linkage Disequilibrium



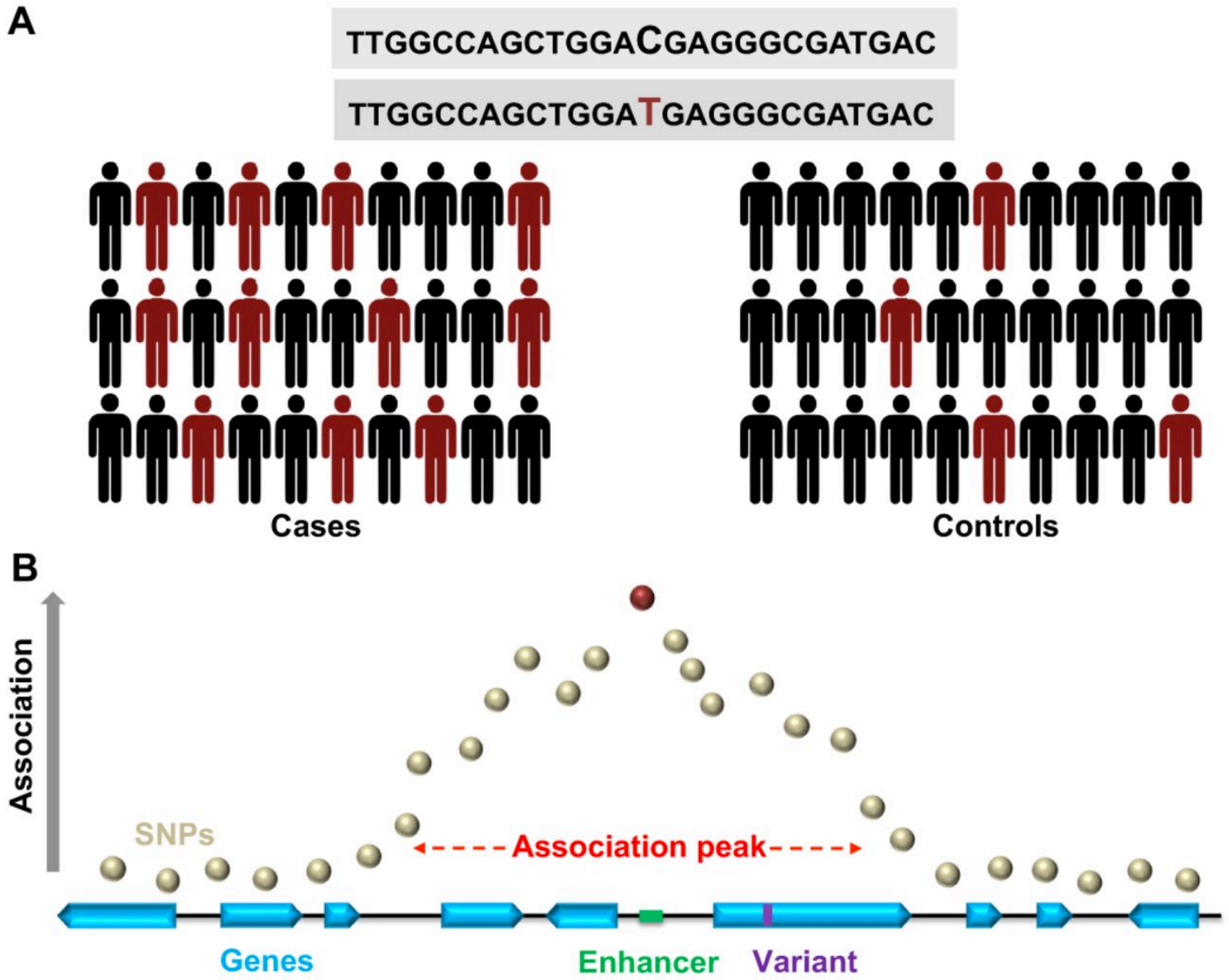


- Linkage: the nonrandom association of alleles at different loci.
- Mutation, admixture, founder effect, random genetic drift, and selection may all lead to an initial association.
- Recombination breaks linkage

Outline


Introduction
GWAS
Population Stratification
Methods to correct
Further Challenges

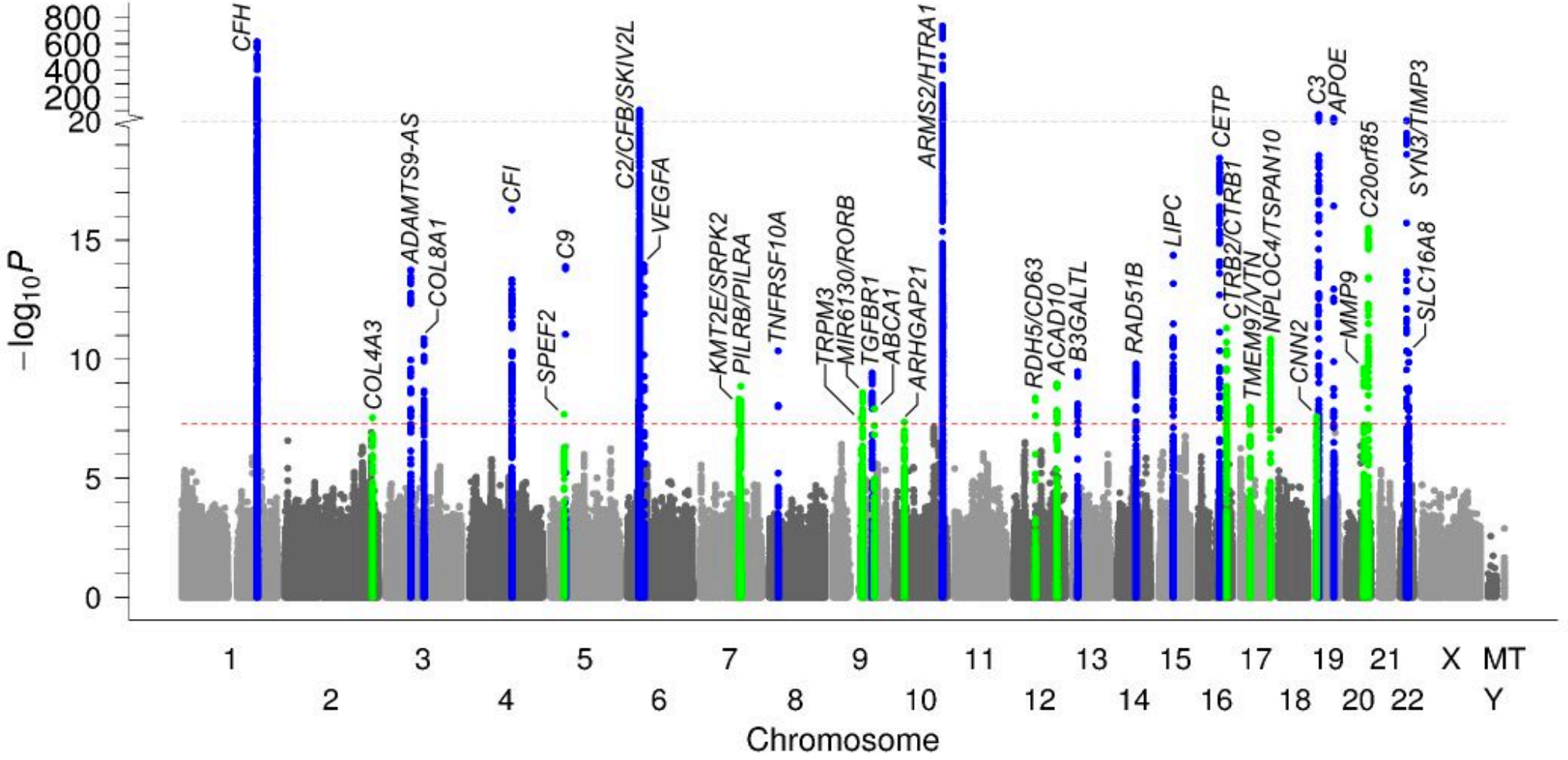
Genome-Wide Association Study (GWAS)



Ref: Wangler, Michael F., Yanhui Hu, and Joshua M. Shulman. "Drosophila and genome-wide association studies: a review and resource for the functional dissection of human complex traits." *Disease Models & Mechanisms* 10.2 (2017): 77-88.

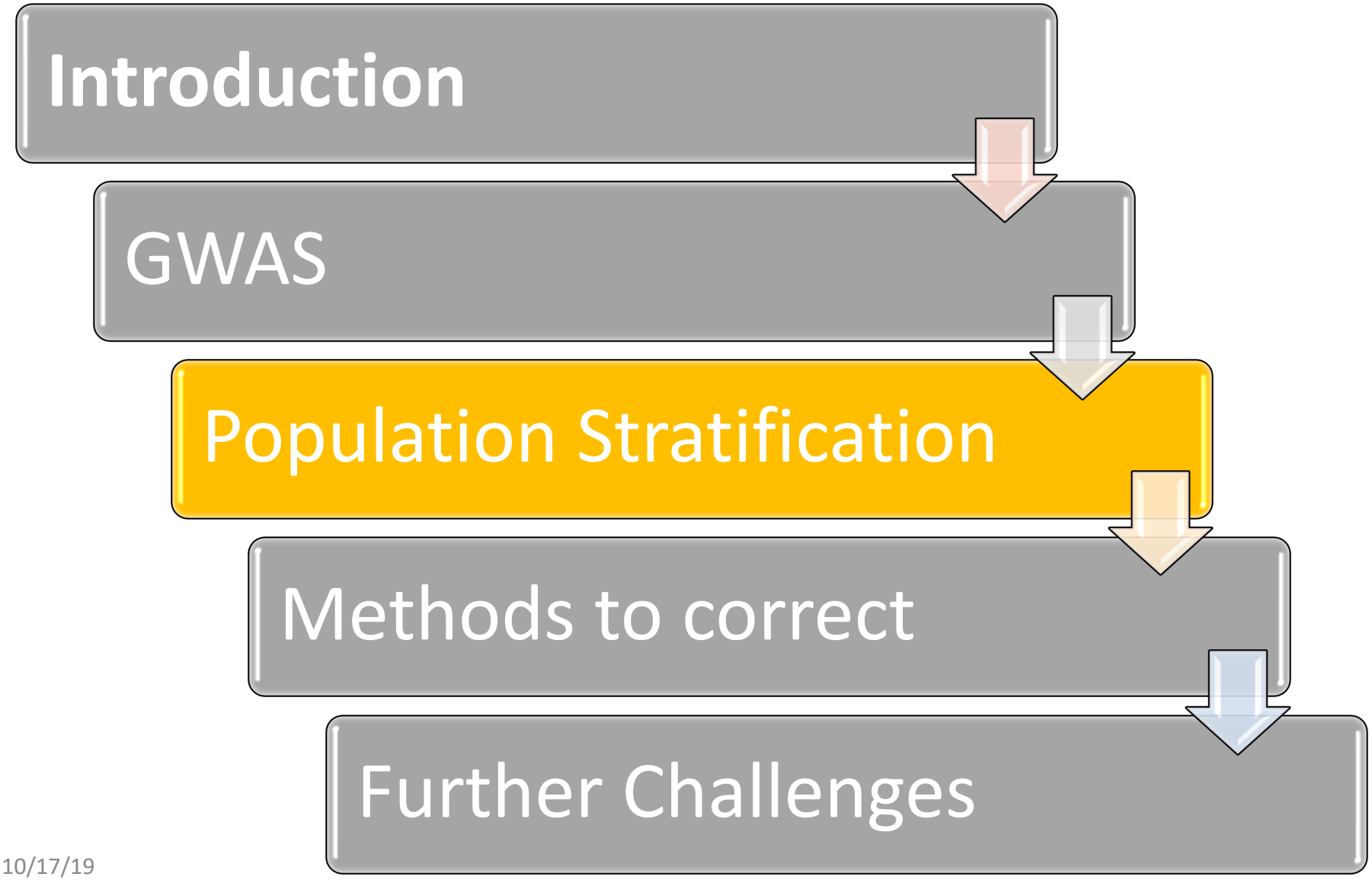
GWAS results – Manhattan plot

https://www.genenames.org/



Ref: Fritsche, Lars G., et al. "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants." *Nature genetics*48.2 (2016): 134.

Outline


Introduction
GWAS
Population Stratification
Methods to correct
Further Challenges

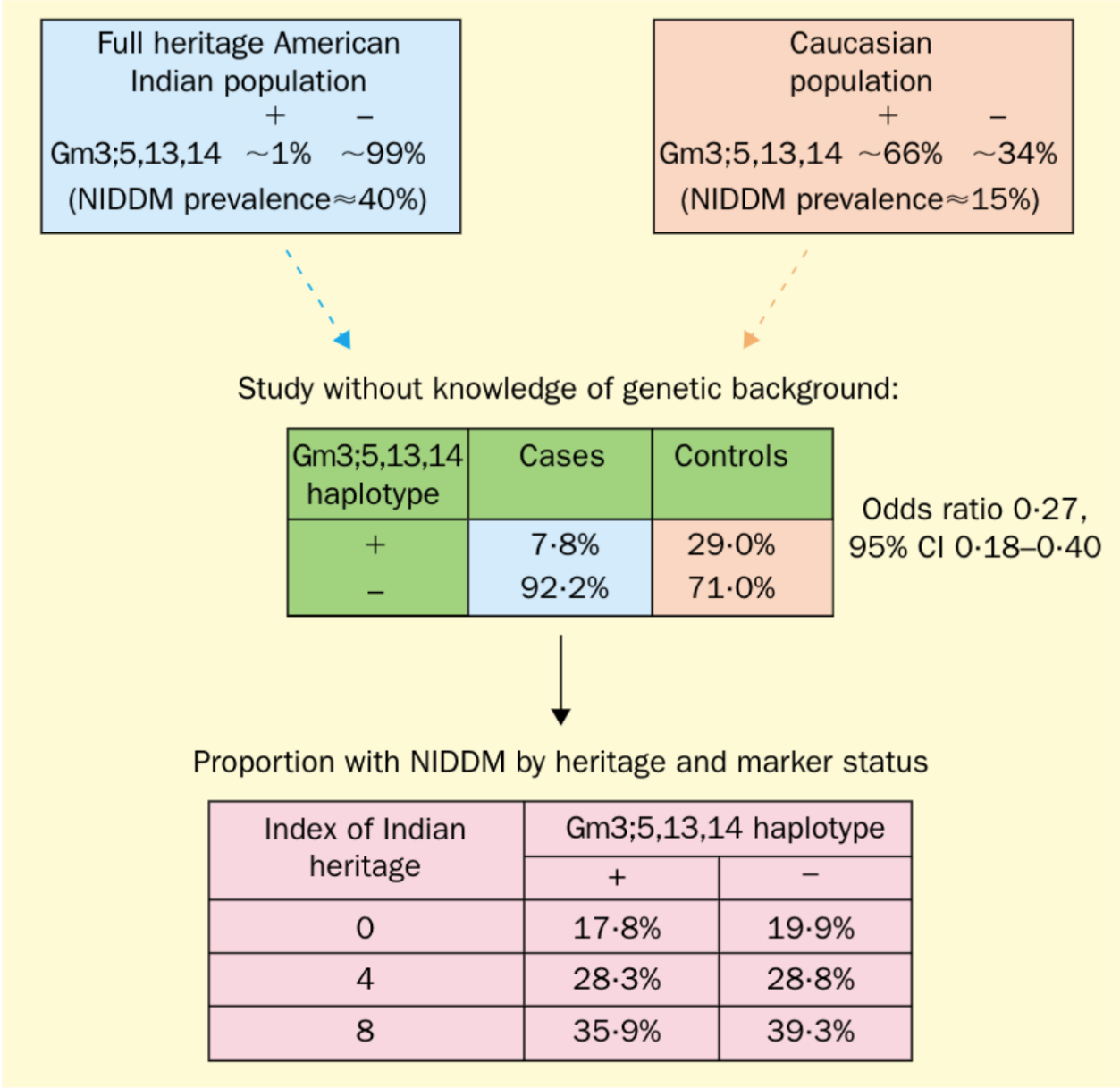
The discovery of the ‘successful-use-of-selected-hand-instruments gene’ (SUSHI).

Beware the chopsticks gene

D Hamer ✉ & L Sirota

Molecular Psychiatry **5**, 11–13 (2000) | Download Citation ↓

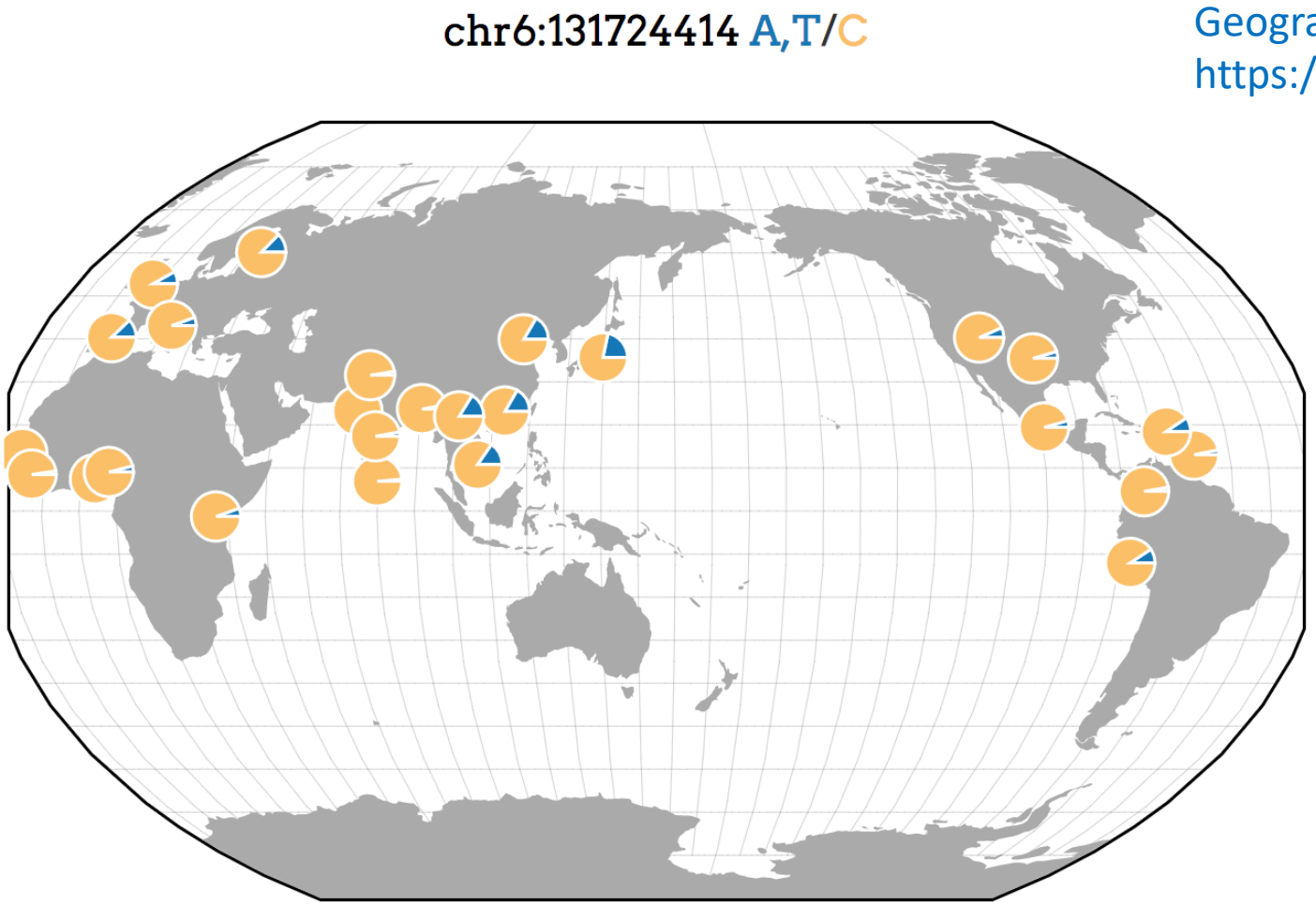
Spurious Association due to Population Stratification



- Systematic differences in allele frequencies due to the differences in sample ancestries.
- Often-cited reason for lack of replication.
- Can lead to both false positive or false negative findings.

Ref: Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, *361*(9357), 598-604.

Geography of Genetic Variants



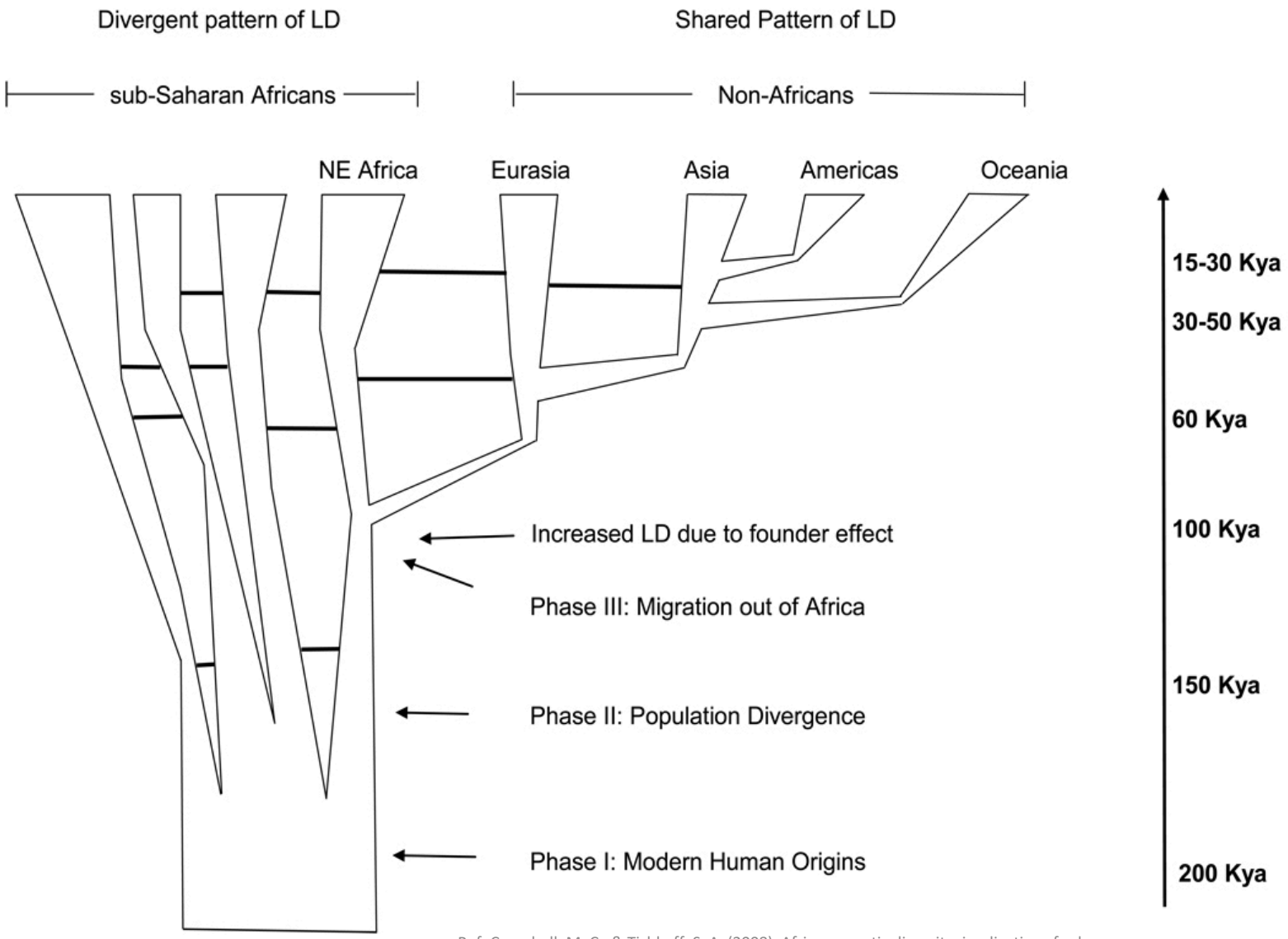
Geography of Genetic Variants Browser
https://popgen.uchicago.edu/ggv

Ref: Cardon, Lon R., and Lyle J. Palmer. "Population stratification and spurious allelic association." *The Lancet* 361.9357 (2003): 598-604.
Gelernter, Joel, et al. "No association between an allele at the D2 dopamine receptor gene (DRD2) and alcoholism." *Jama*266.13 (1991): 1801-1807.

Human migration out of Africa

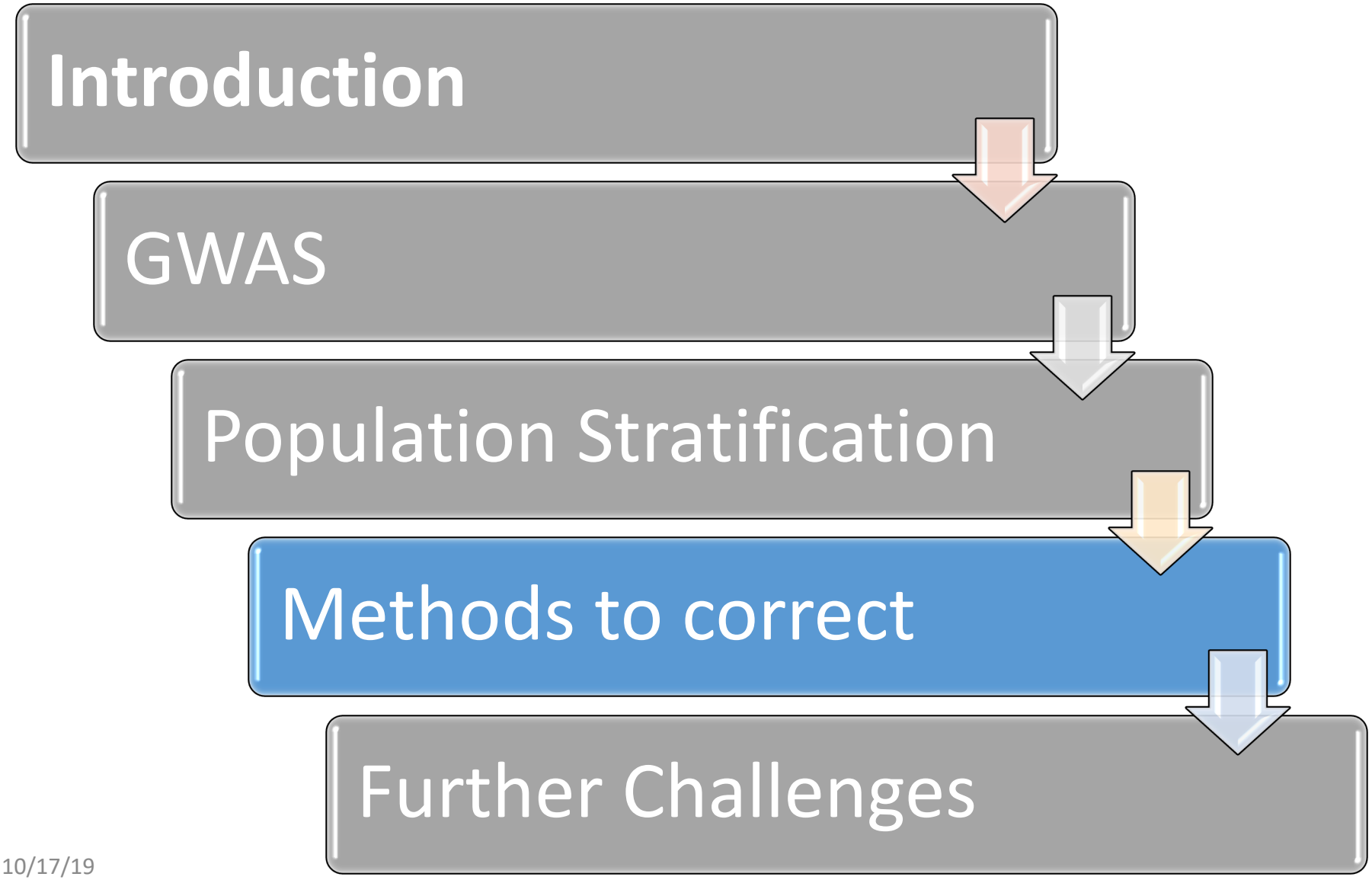
Ref: Demenocal, P. B., & Stringer, C. (2016). Human migration: Climate and the peopling of the world. *Nature*, *538*(7623), 49.

African Genetic Diversity

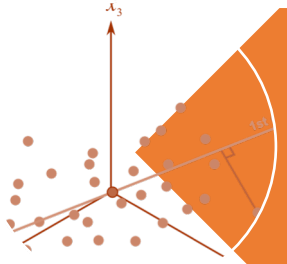


Ref: Campbell, M. C., & Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.*, *9*, 403-433.

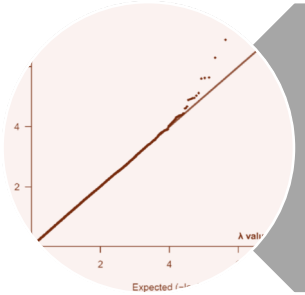
Outline


Introduction
GWAS
Population Stratification
Methods to correct
Further Challenges

Methods to correct population stratification:



Principal Component Analysis (PCA)

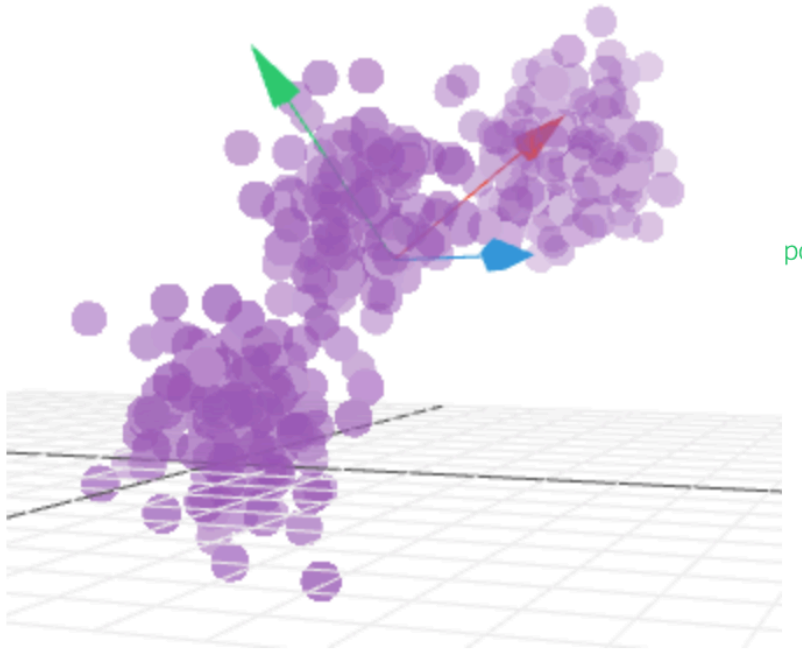


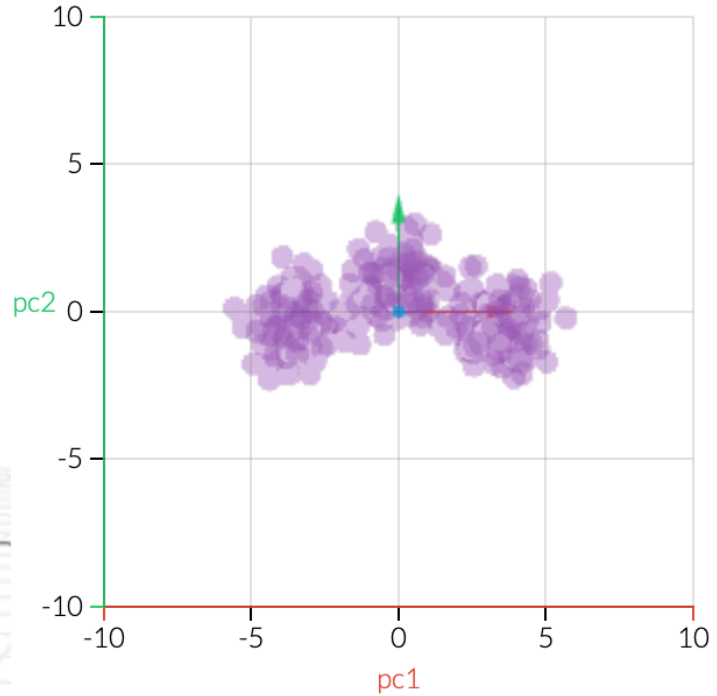
Genomic Control

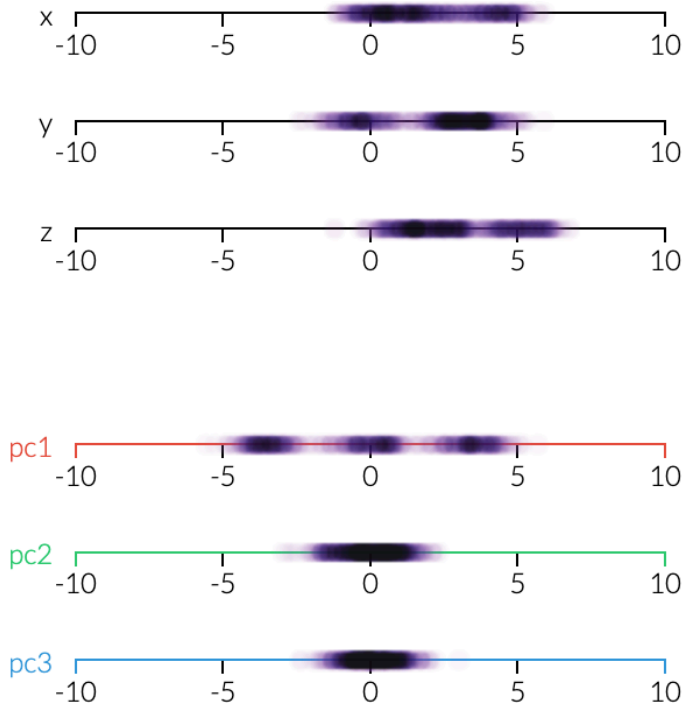
Family-based study design

Unsupervised Learning - PCA

- A dimension reduction method widely used in GWAS
- Principal Components (PCs): Viewed as continuous axes of variation that reflect genetic variation due to ancestry
- Identifies several top PCs and uses them as covariates in the association analyses to control for population stratification.

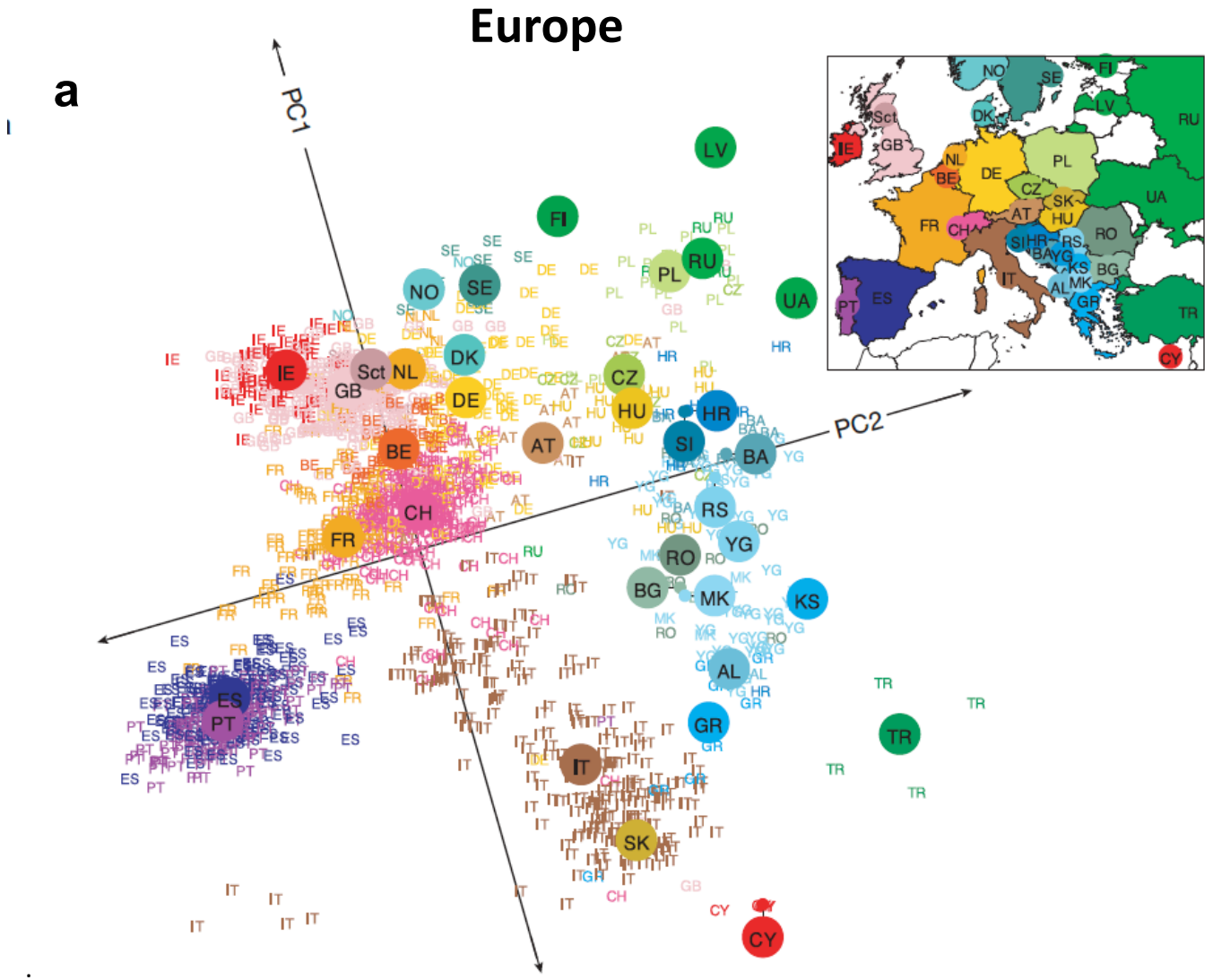


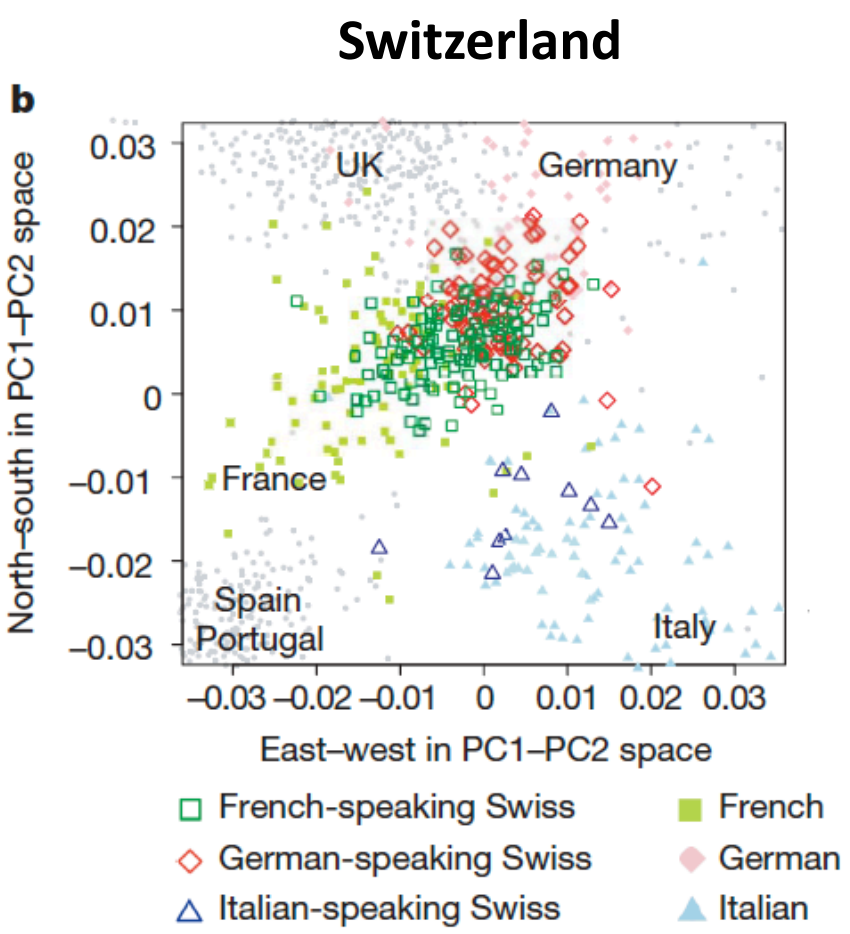




Great Visualization here!
http://setosa.io/ev/principal-component-analysis/

PCA for identifying and adjusting ancestry

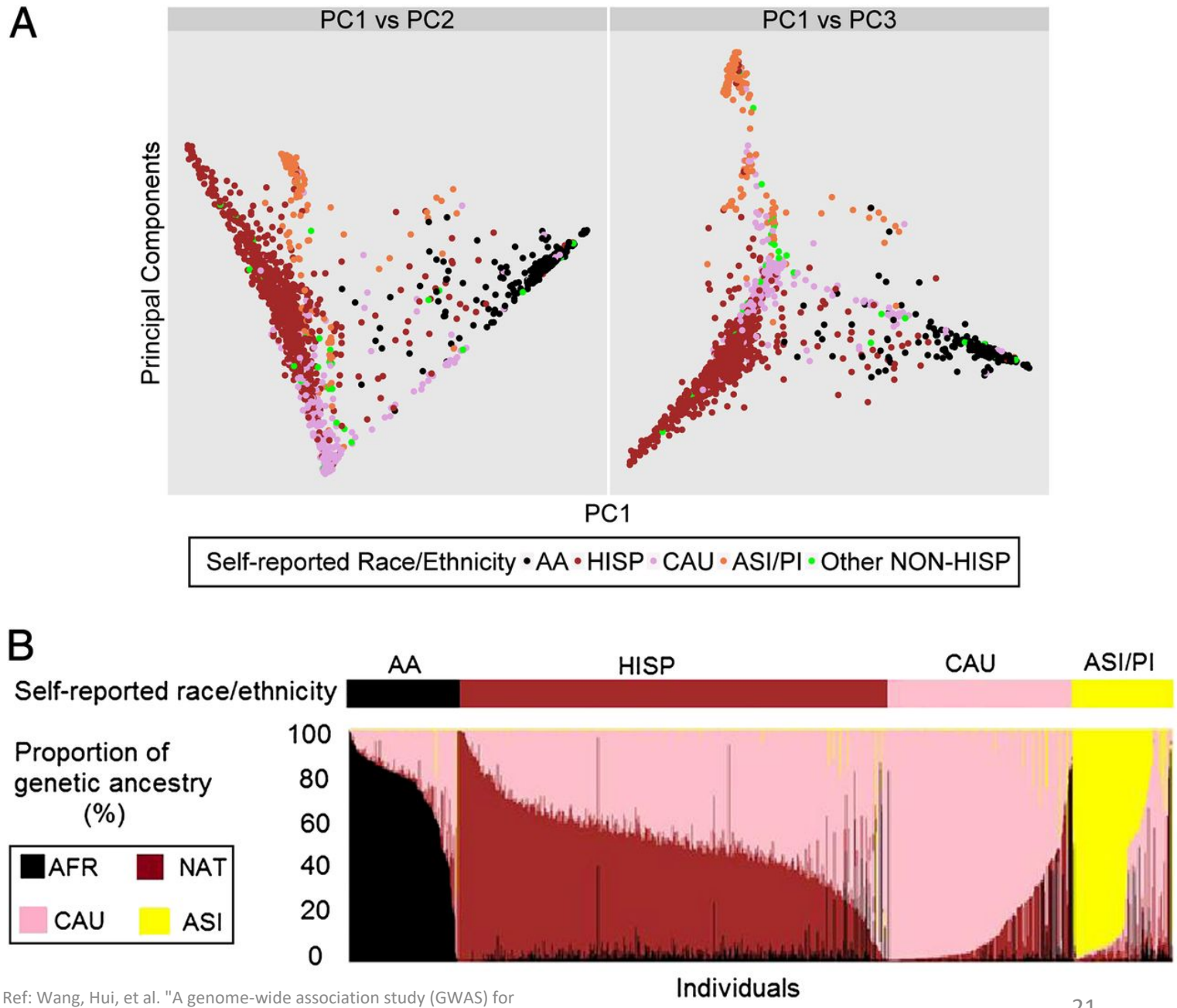




Ref: Novembre, John, et al. "Genes mirror geography within Europe." *Nature* 456.7218 (2008): 98

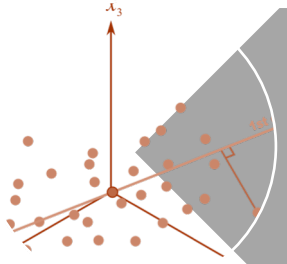
Genetic Ancestry vs Self-reported Race/ethnicity

Is race biological or socially constructed?

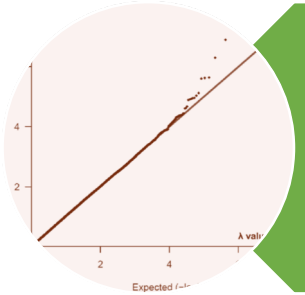


Ref: Wang, Hui, et al. "A genome-wide association study (GWAS) for bronchopulmonary dysplasia." *Pediatrics* 132.2 (2013): 290-297.

Methods to correct population stratification:



Principal Component Analysis (PCA)



Genomic Control

Family-based study design

Genomic Control

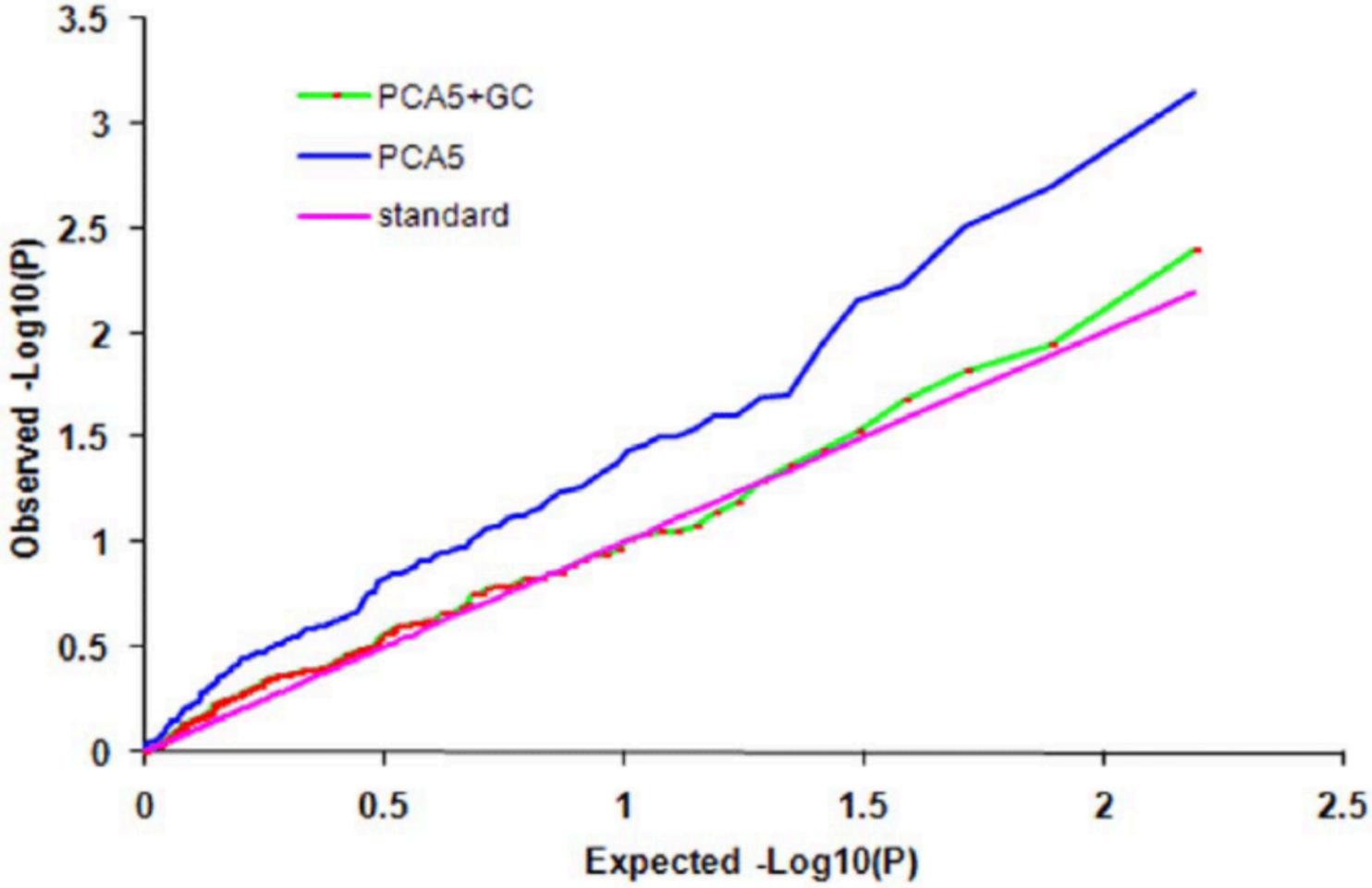
- Based on Devlin and Roeder (1999)'s assumption that the overdispersion of chi-square test statistics caused by population stratification is roughly constant across the genome.
- To estimate λ, two choices are natural: a robust estimator such as the median of the chi-square test statistics, divided by 0.456 (Devlin and Roeder 1999), or the mean (Reich and Goldstein 2001).
- Modifies the association test statistic by a common factor λ for all SNPs.

$$\chi^2_{GC} = \chi^2_{RAW}/\lambda$$

Ref: Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997-1004.
Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, *20*(1), 4-16.

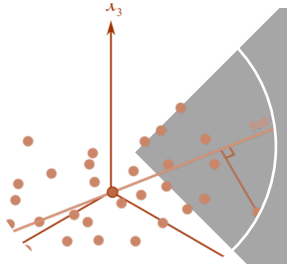
Genomic Control combined with PCA

The cumulative distributions of observed -Log10(P) values before and after genomic control (GC) in PCA5 model.

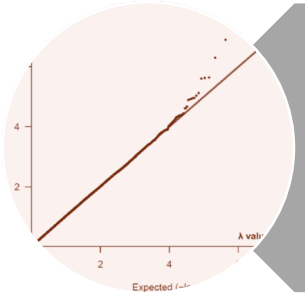


Ref: Jia, Limeng, et al. "Allelic analysis of sheath blight resistance with association mapping in rice." *PLoS One* 7.3 (2012): e32703.

Methods to correct population stratification:



Principal Component Analysis (PCA)



Genomic Control

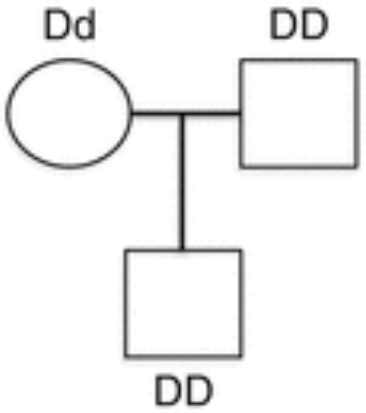
Family-based study design

Family-based study design

- Use probands and their relatives, typically either parents or siblings.
- Robust against confounding due to population substructures
- Can test both linkage and association
- Family-Based Association Test (FBAT): relies solely on within-family information by constructing a score test that essentially provides a correlation between phenotype and genotype.
- Measured Genotype Approach(Mixed models): a variance components framework utilizes a mixed model in which familial relationships are accounted for using random effects and genetic variants are incorporated as fixed effects.

Trio-design -Transmission disequilibrium test

Transmission counts for one family



	D non transmitted	d non transmitted
D transmitted	0	1
d transmitted	0	0

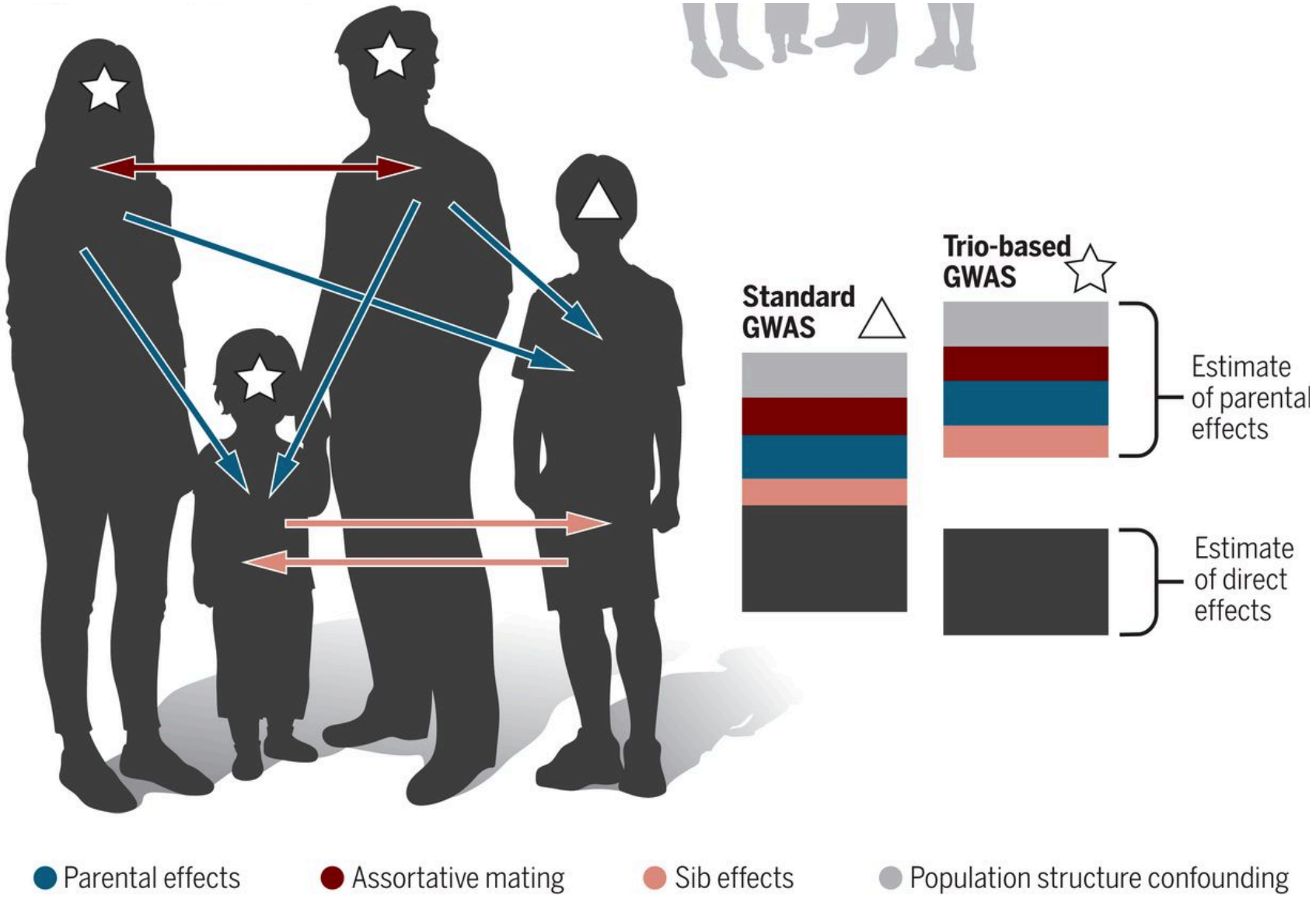
TDT in a sample of n=90 families with heterozygote parents

	D non transmitted	d non transmitted	Total
D transmitted	a = 0	b = 120	120
d transmitted	c = 60	d = 0	60
Total	60	120	2n=180

The TDT tests for the null hypothesis of Mendelian allelic transmission D:d=1:1

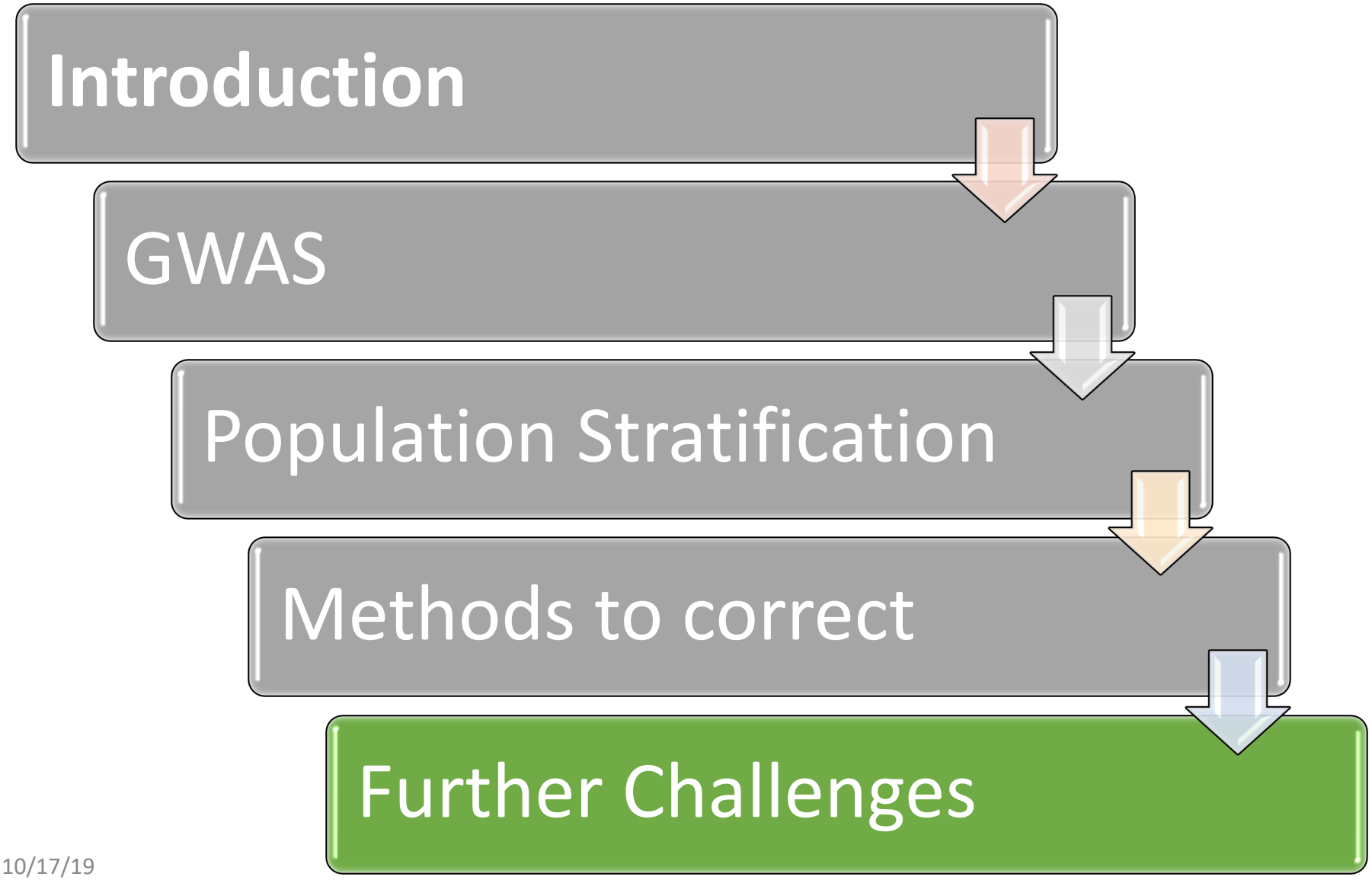
Null hypothesis $H_0 : \frac{b}{b+c} = \frac{c}{b+c} = 0.5$ $\chi^2 = \frac{(b-c)^2}{b+c} = 20$ P-value is 7.7×10^{-6}

Trio-based GWAS

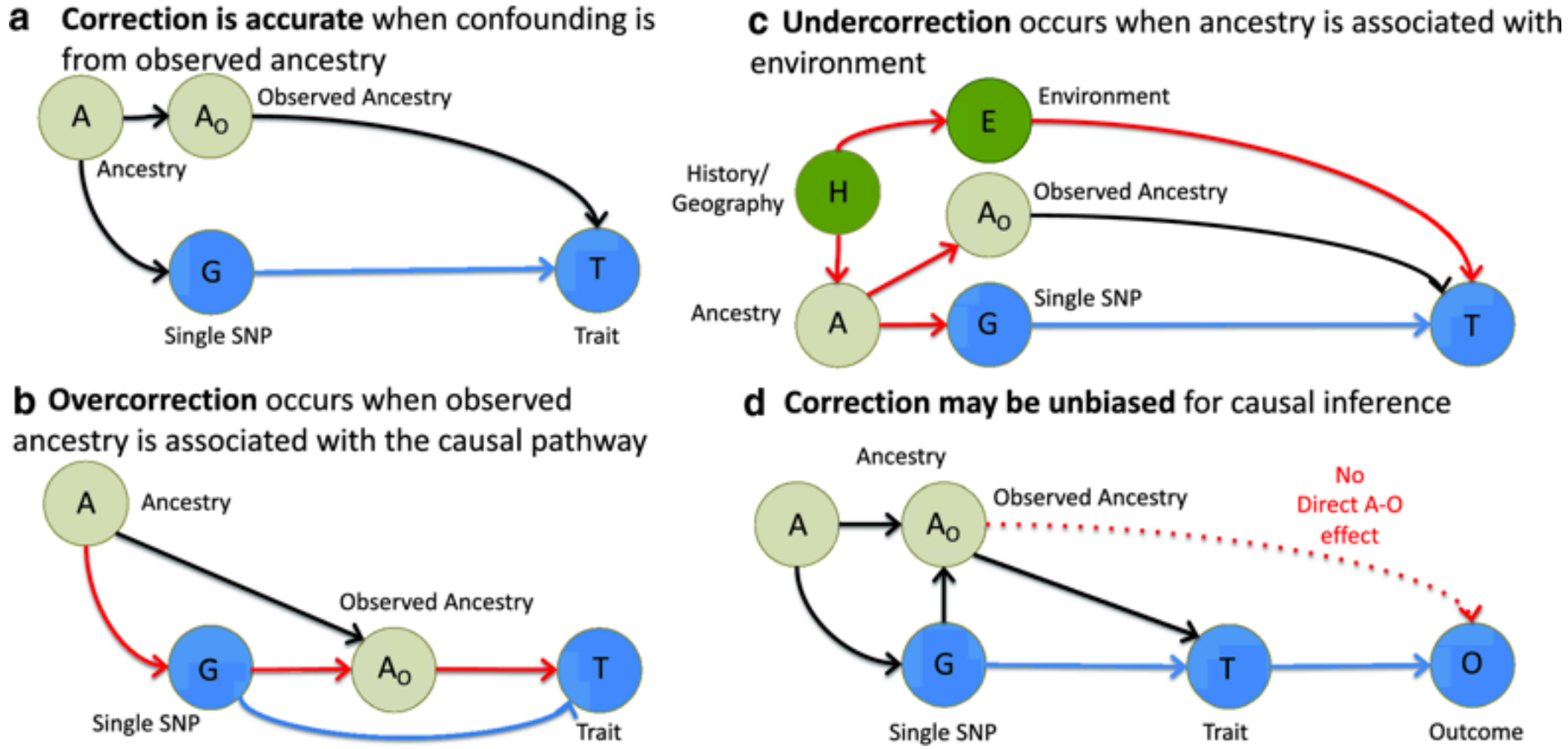


Ref: Young, Alexander I., et al. "Deconstructing the sources of genotype-phenotype associations in humans." *Science* 365.6460 (2019): 1396-1400.

Outline

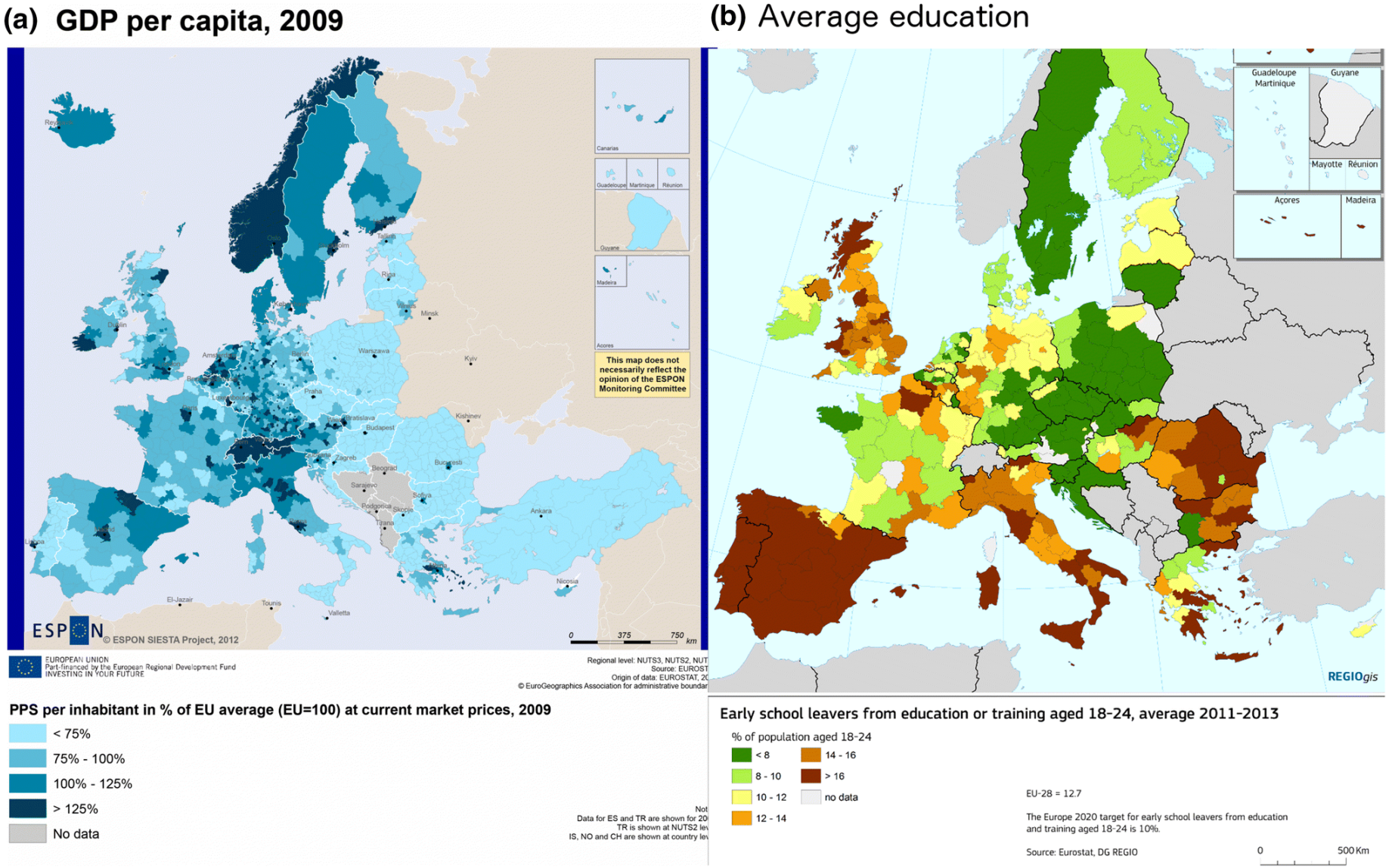

Introduction
GWAS
Population Stratification
Methods to correct
Further Challenges

Causal Inference Directed Acyclic Graph



Ref: Lawson, Daniel John, et al. "Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?." *Human genetics* (2019): 1-19.

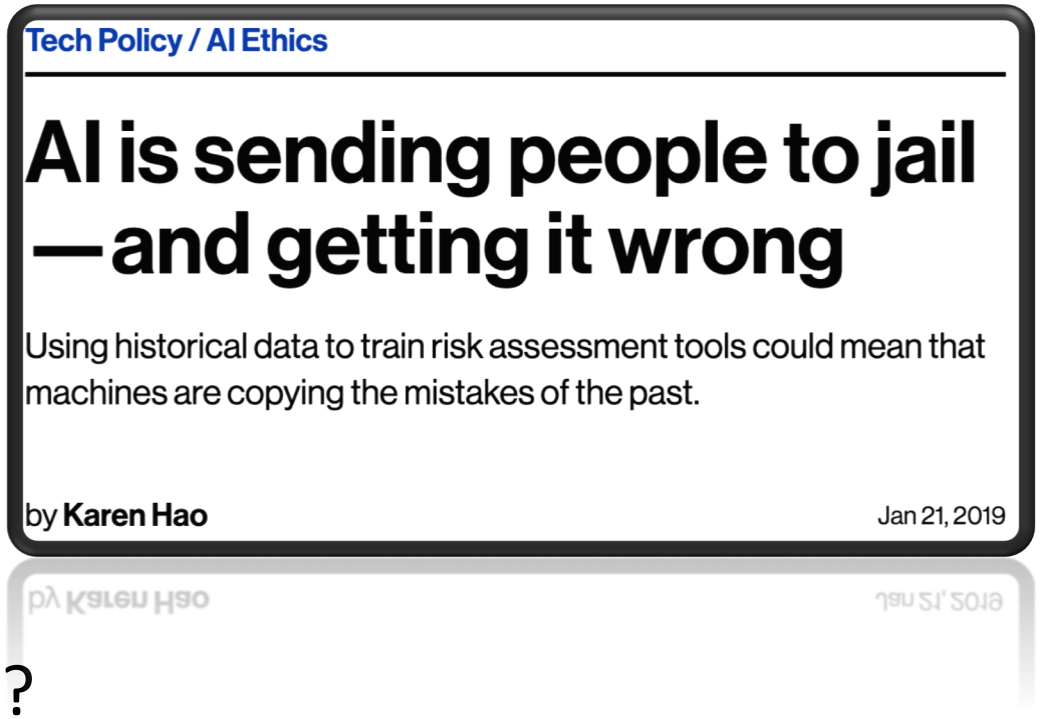
Education correlates with GDP – confounding for Education GWAS?



Ref: Lawson, Daniel John, et al, 2019

Concerns for policy implications of GWAS

- Association versus causation
- Undercorrection versus overcorrection of population stratification
- Nature versus nurture
- Lack of diversity in data collection
- The usage of race in current pharmacogenomics studies
- More regulations beyond The Genetic Information Nondiscrimination Act of 2008?


Tech Policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** Jan 21, 2019

Q & A

Thank you!