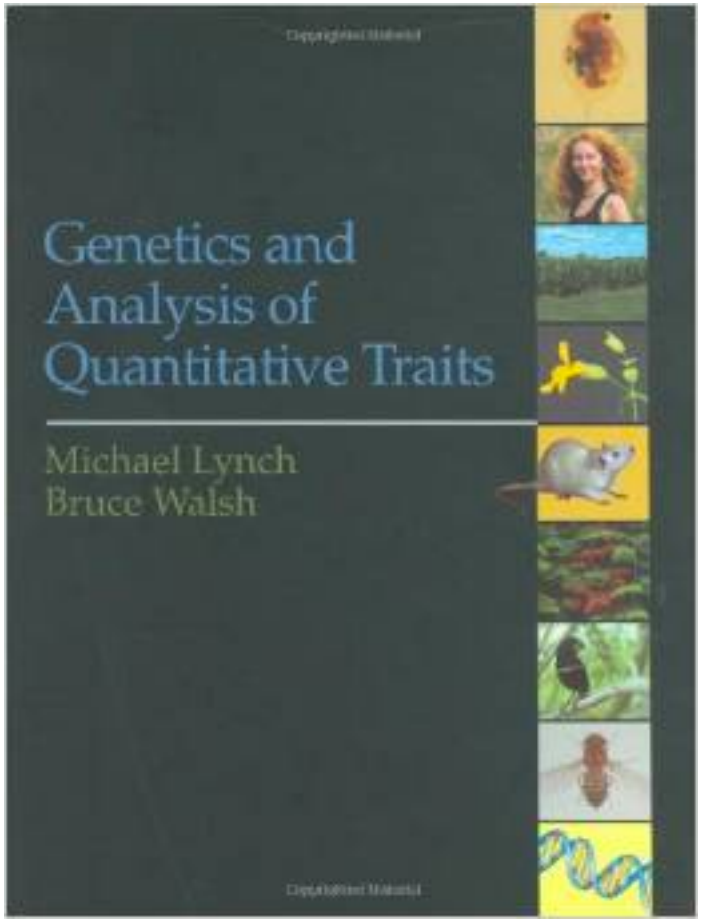
Survey of the Principals of Quantitative/Medical Genetics



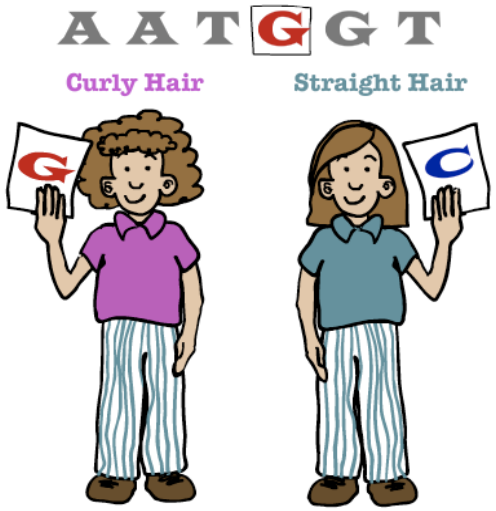
http://nitro.biosci.arizona.edu/zbook/book.html

What is Medical Genetics?

Medical genetics is the study of the relationship between genetic variation and human disease.

Genetic variants cause differences in phenotypes

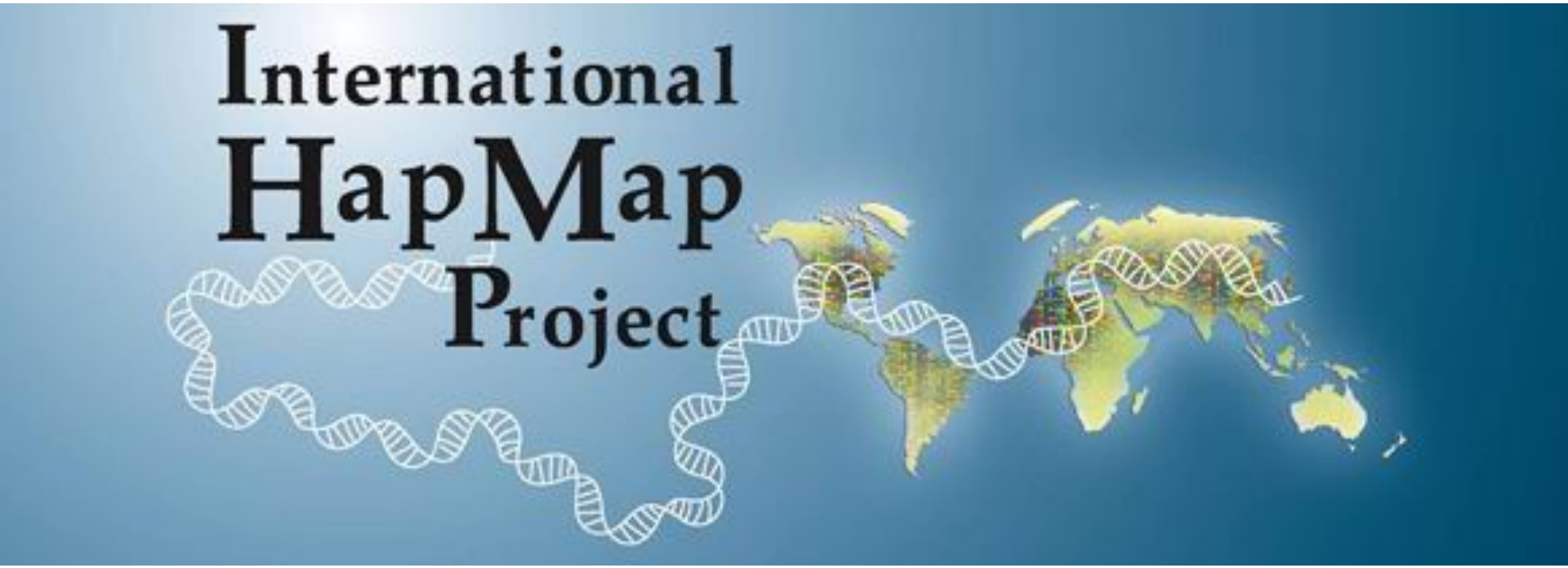
- Without genetic variation we would all be identical twins (like the Cavendish bananas in the supermarket).
- For example, some SNP may cause people to have different hair



www.23andme.com/gen101/snps

What more powerful form of study of mankind could there be than to read our own instruction book?

-Francis Collins, June 2000



The International HapMap Project is a multi-country effort to identify and catalog genetic similarities and differences in human beings. Using the information in the HapMap, researchers will be able to find genes that affect health, disease, and individual responses to medications and environmental factors.
October 2002

Association study

cases: people with a disease

SNP X in DNA

```
Case 1:   A
Case 2:   A
Case 3:   A
Case 4:   A
Case 5:   A
….
Case 1000000: A
```

controls: people w/o a disease

```
Control 1:   T
Control 2:   T
Control 3:   T
Control 4:   T
Control 5:   T
….
Control 1000000: T
```

- This SNP is correlated with or “associated” with a disease
- People who have “A” allele are more likely to have the disease

More realistic association study

cases (+)

400 cases have A
600 cases have T

$\hat{p}_X^+ = 0.4$

controls (-)

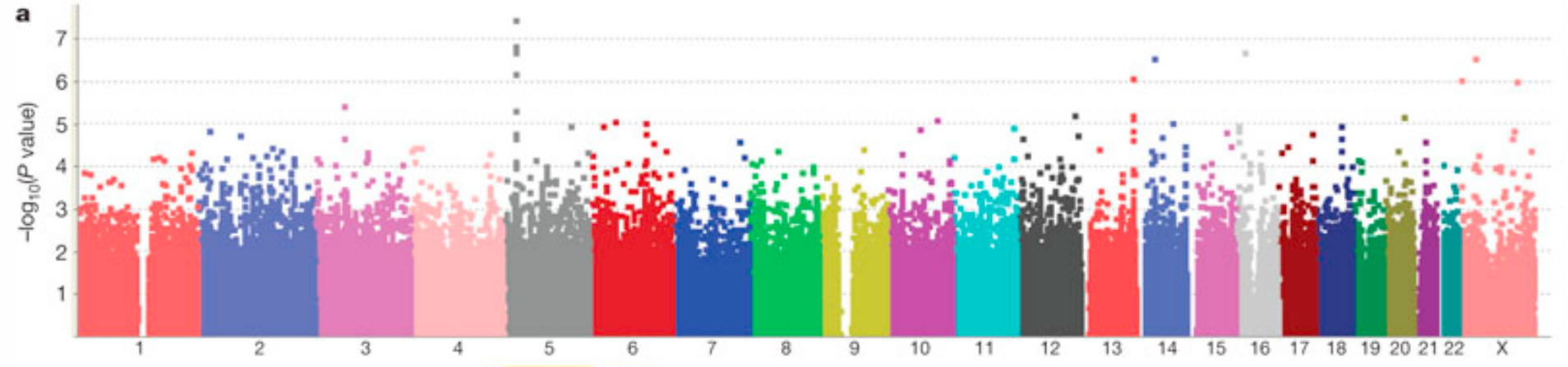
300 controls have A
700 controls have T

$\hat{p}_X^- = 0.3$

- We compute correlation (association statistic) between SNP and a disease
  - Association statistic is based on allele freq. difference ($\hat{p}_X^+ - \hat{p}_X^-$)
  - The larger the difference, the higher the correlation
- If correlation is above certain threshold, SNP is associated with a disease

Genome-wide Association Studies (GWAS)

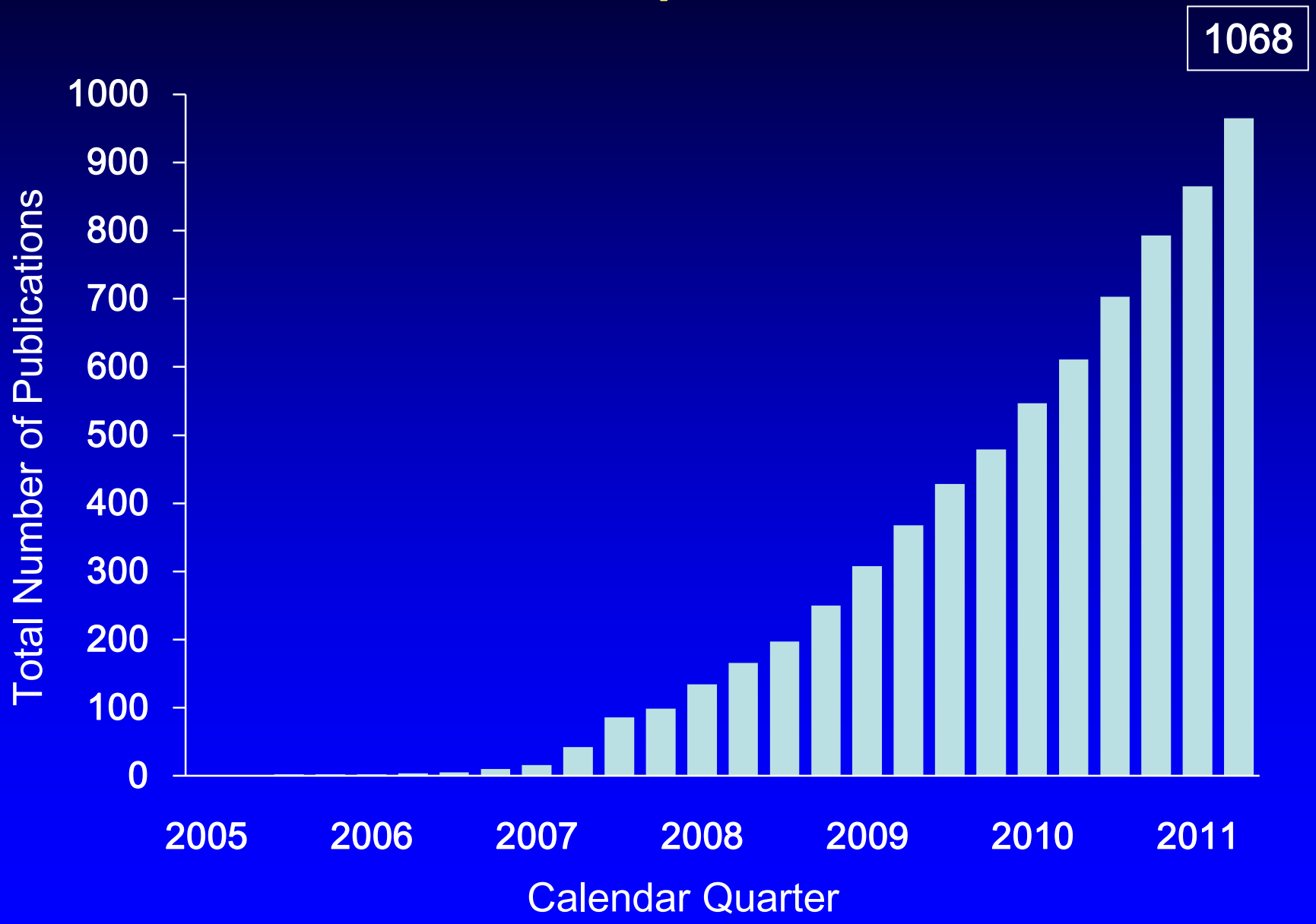
- Collect many SNPs (~1 million) over the whole genome using microarray
- Compute correlation between each SNP and a disease (perform "association test" on each SNP)
- Find SNPs whose correlations are above the threshold



Wang, K., Zhang, H., Ma, D., Bucan, M., et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459, 528-533 (2009).

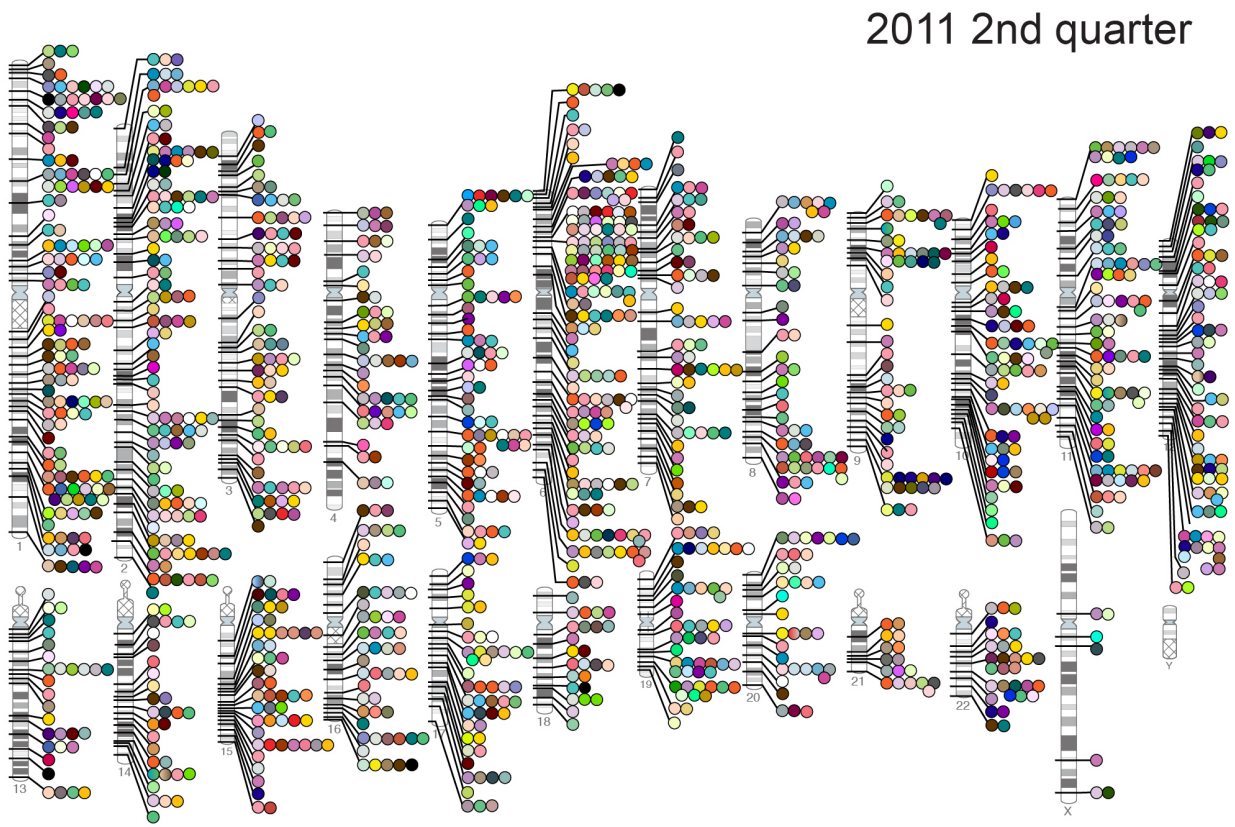
- A peak in the plot means a strong association between SNP and a disease

Published GWA Reports, 2005 – 9/2011



Through 9/30/10 postings

GWAS have identified 1000s of variants associated with 100s of phenotypes



NHGRI GWA Catalog, www.genome.gov/GWAStudies

GWAS Algorithm

Acquire genotype data G and phenotype data Y for N individuals

Quality control genotype data: HW, missingness, relatedness, etc.

For each SNP G_i: test G_i for association with Y accounting for age, sex, and population

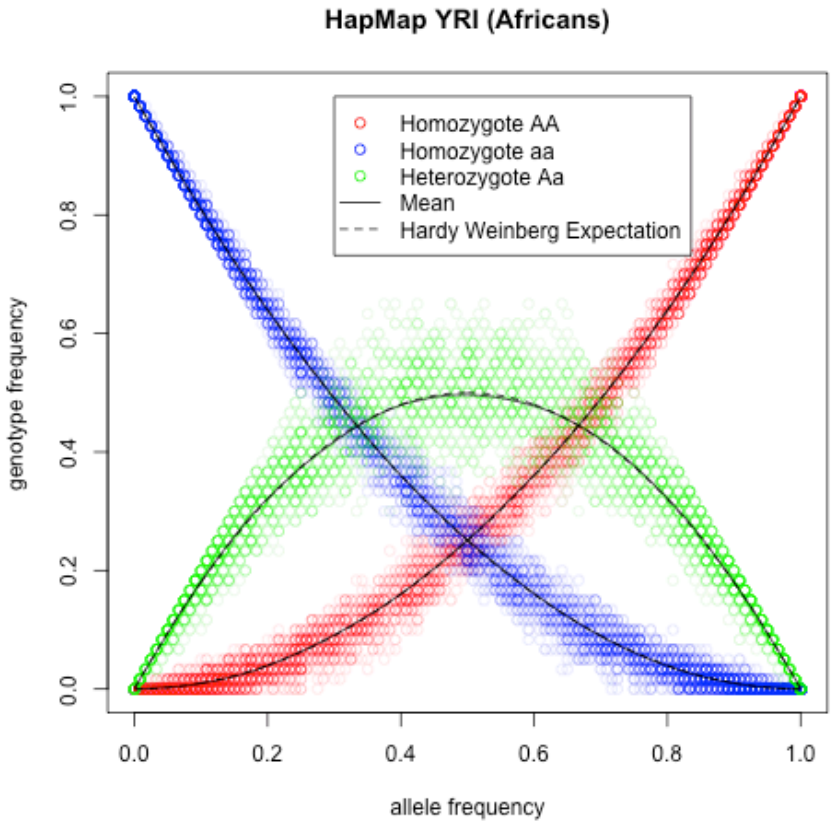
Report all SNPs with p-value < α_{GWAS}

McCarthy et al NRG 2008

Hardy Weinberg

- p=mean(genotypes)/2
- q=1-p
- AA=p^2
- Aa = 2pq
- aa=q^2
- HW=

sum((obs-exp)^2/exp))



Additive Model of Phenotypes

p_i : allele frequency of SNP i

g_{ij} : genotype $\{0,1,2\}$ of individual j at SNP i

$$g_{ij} \sim Bi(p_i) + Bi(p_i)$$

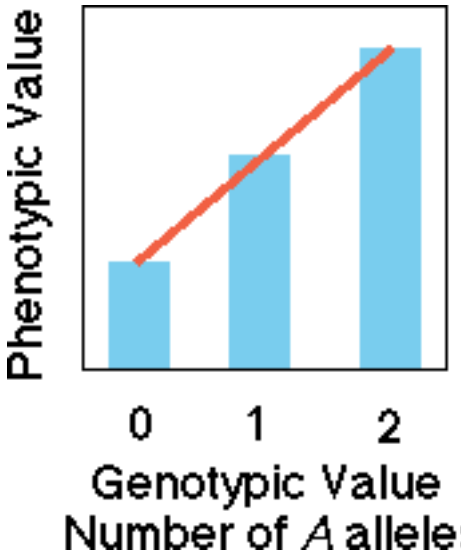
$$\overline{g}_{ij} = \frac{g_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

ε_j : environment of individual j

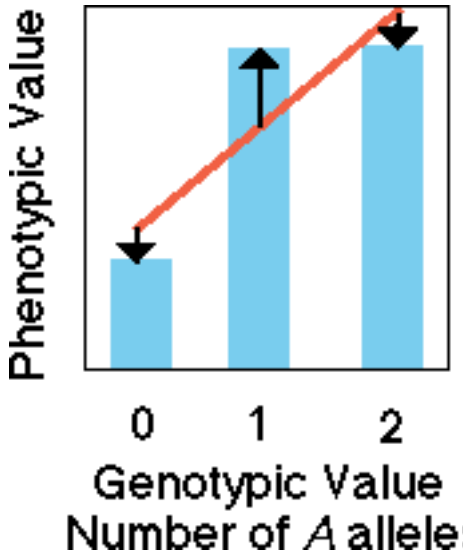
$$\varepsilon_j \sim N(0,\sigma_\varepsilon^2)$$

$$\varphi_j : \text{ phenotype} = \sum_i \beta_i \overline{g}_{ij} + \varepsilon_j \qquad \sigma_A^2 = \sum_i \beta_i^2$$

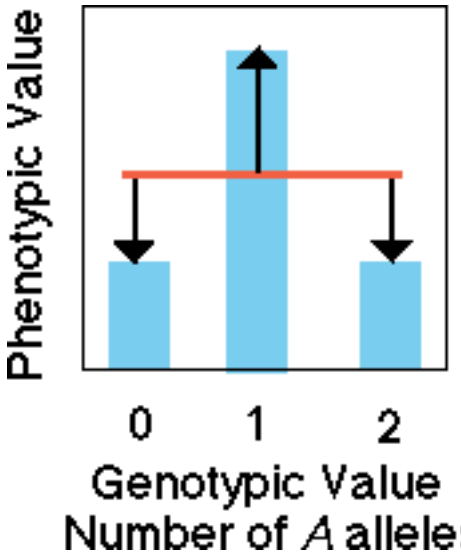
Dominance models of phenotype



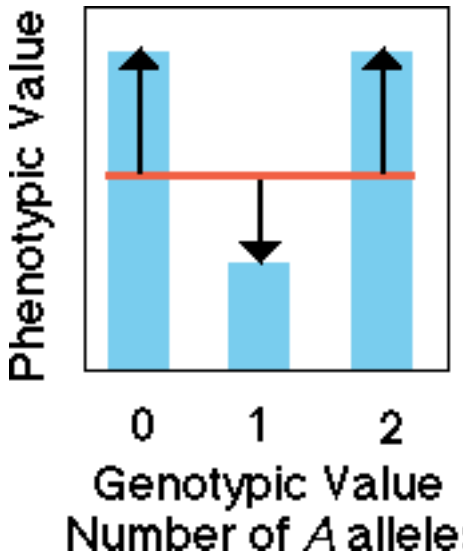
Additive
Co-dominant



Dominant



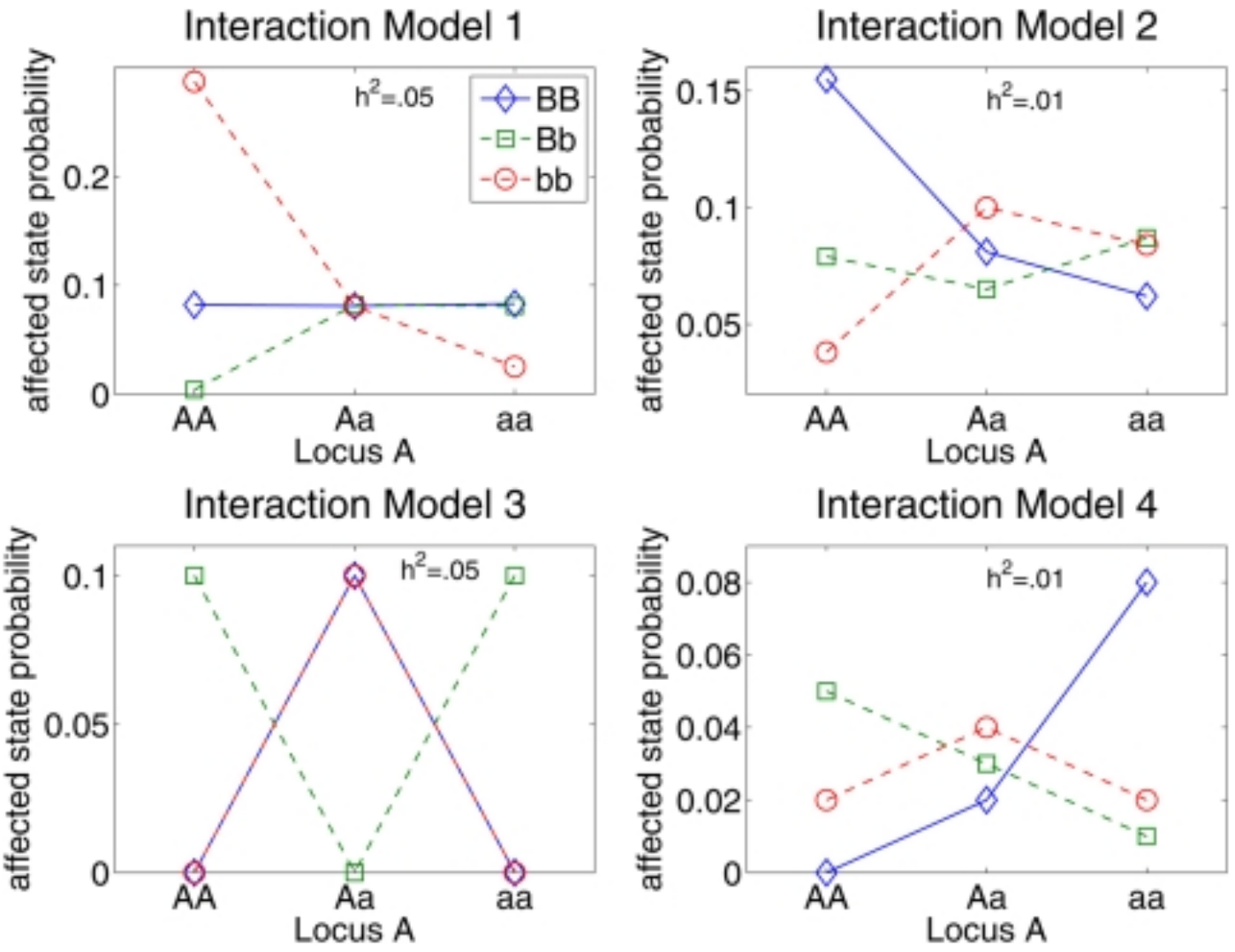
Overdominant



Underdominant

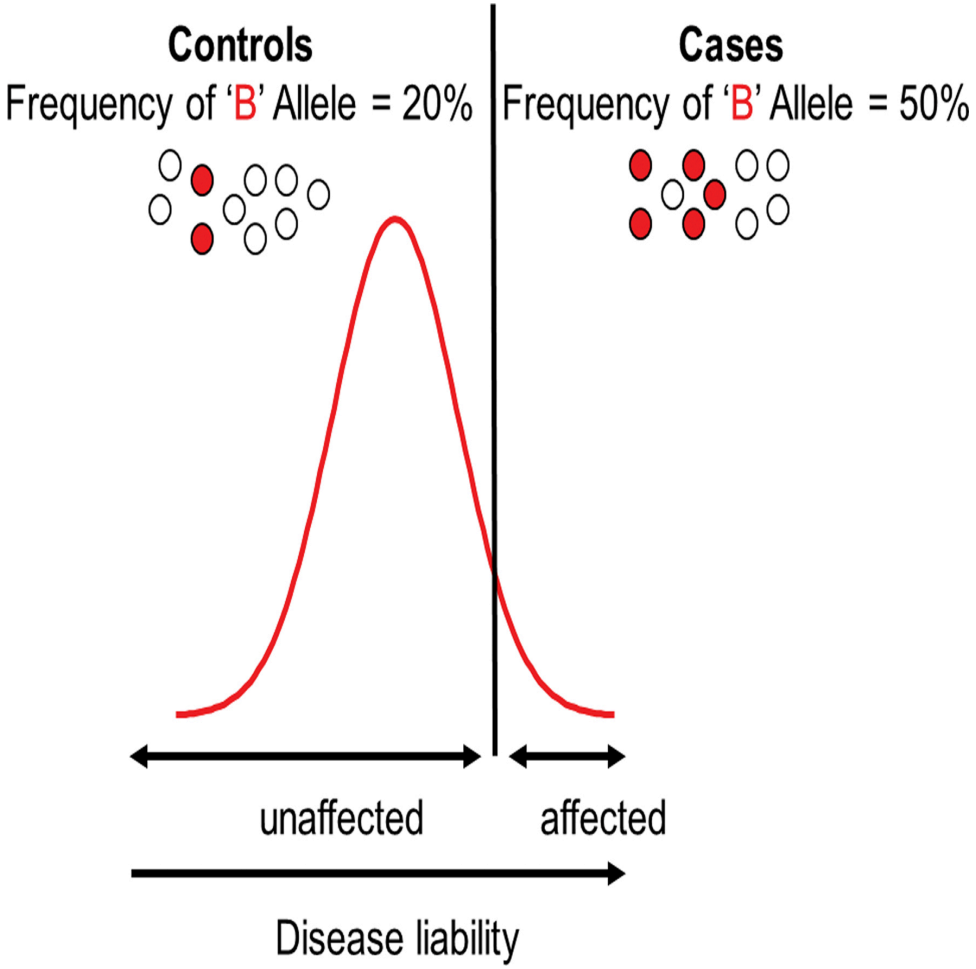
Barry Sinervo @USC

Epistatic Models of Phenotype



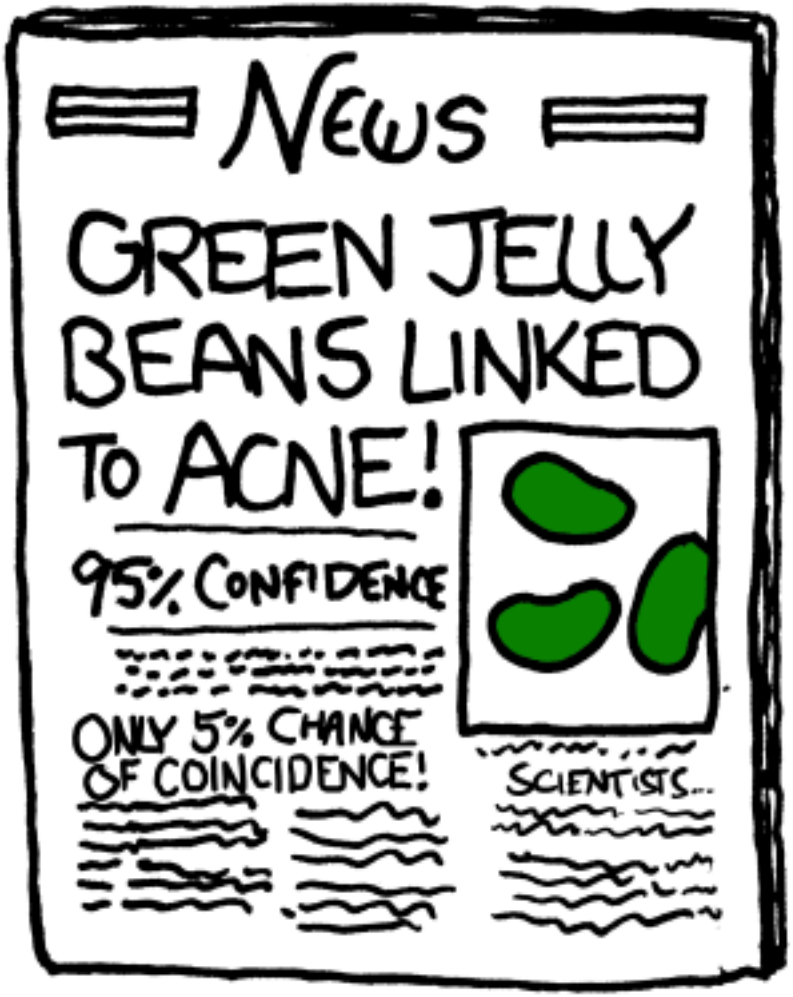
McKinney et al, 2009 PLoS Genetics

Liability Threshold Model of Phenotype

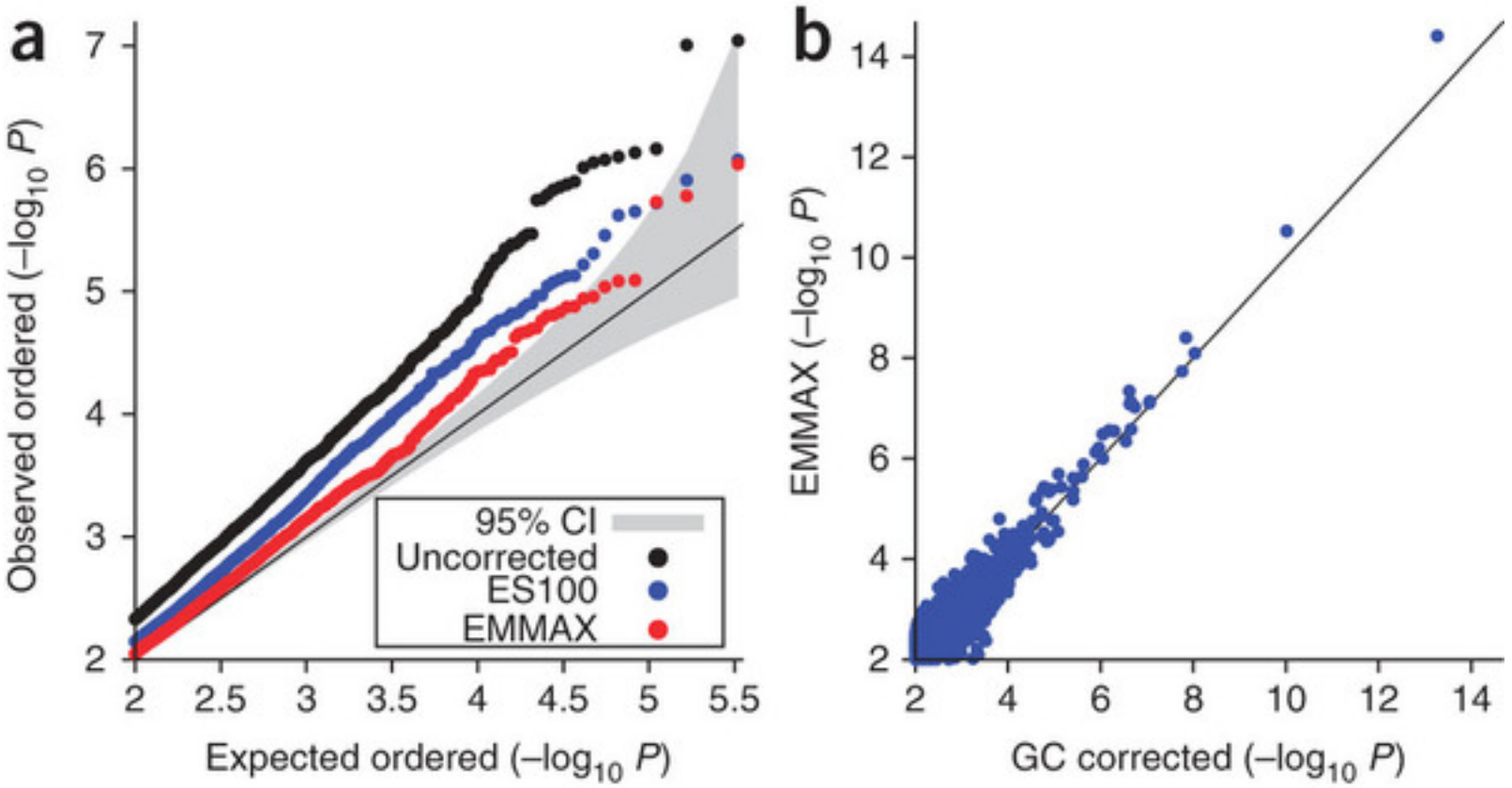


Multiple Hypothesis Testing

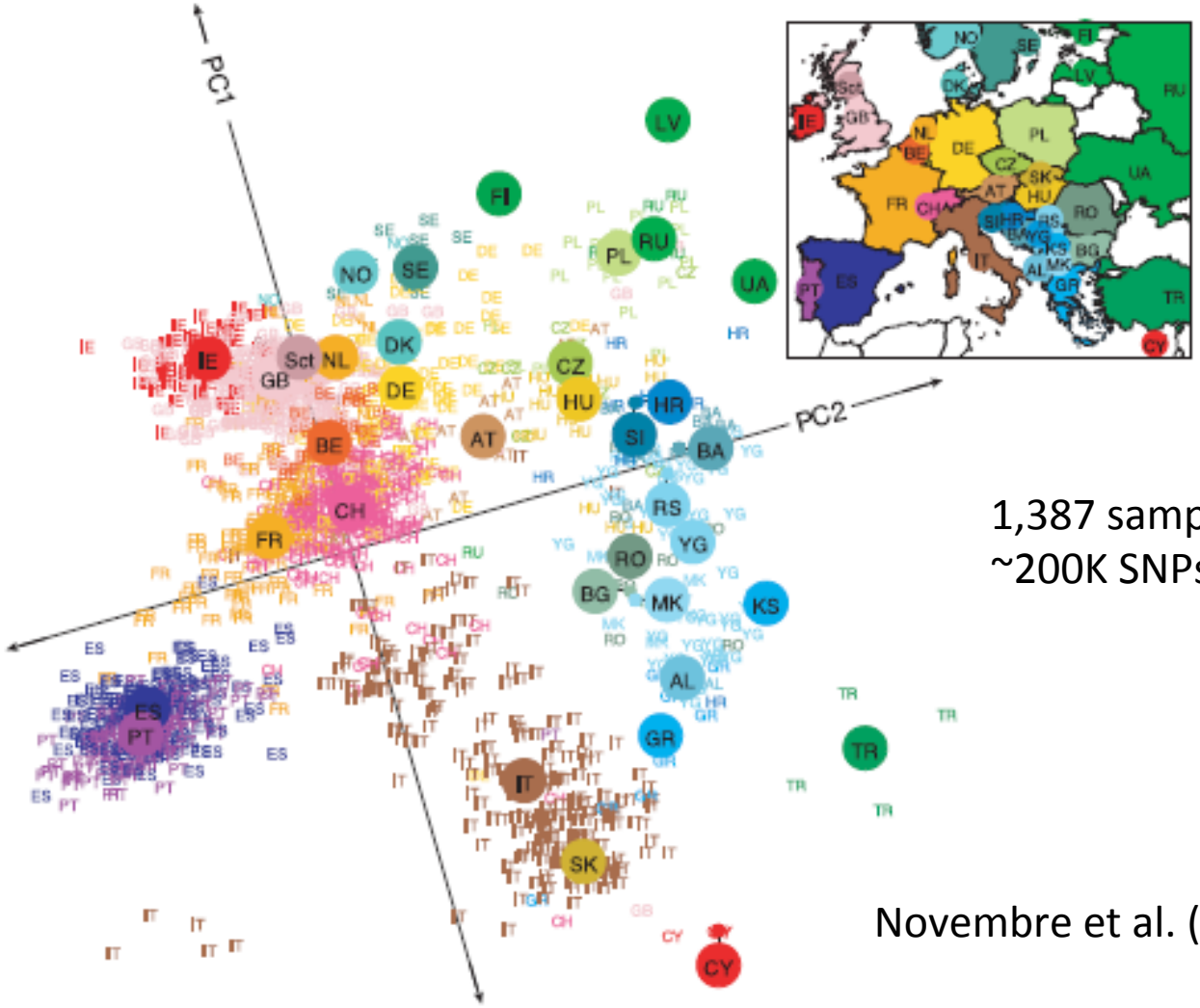
- If we make a single test we use p-value < 0.05
- What if we test 1,000,000 things?
- What if they are not independent?



QQ plots provide a mechanism to identify inflation



Population structure can cause inflation and PCA can correct this inflation



1,387 samples
~200K SNPs

Novembre et al. (2008).

PCA

- Input Kinship (Covariance) Matrix:

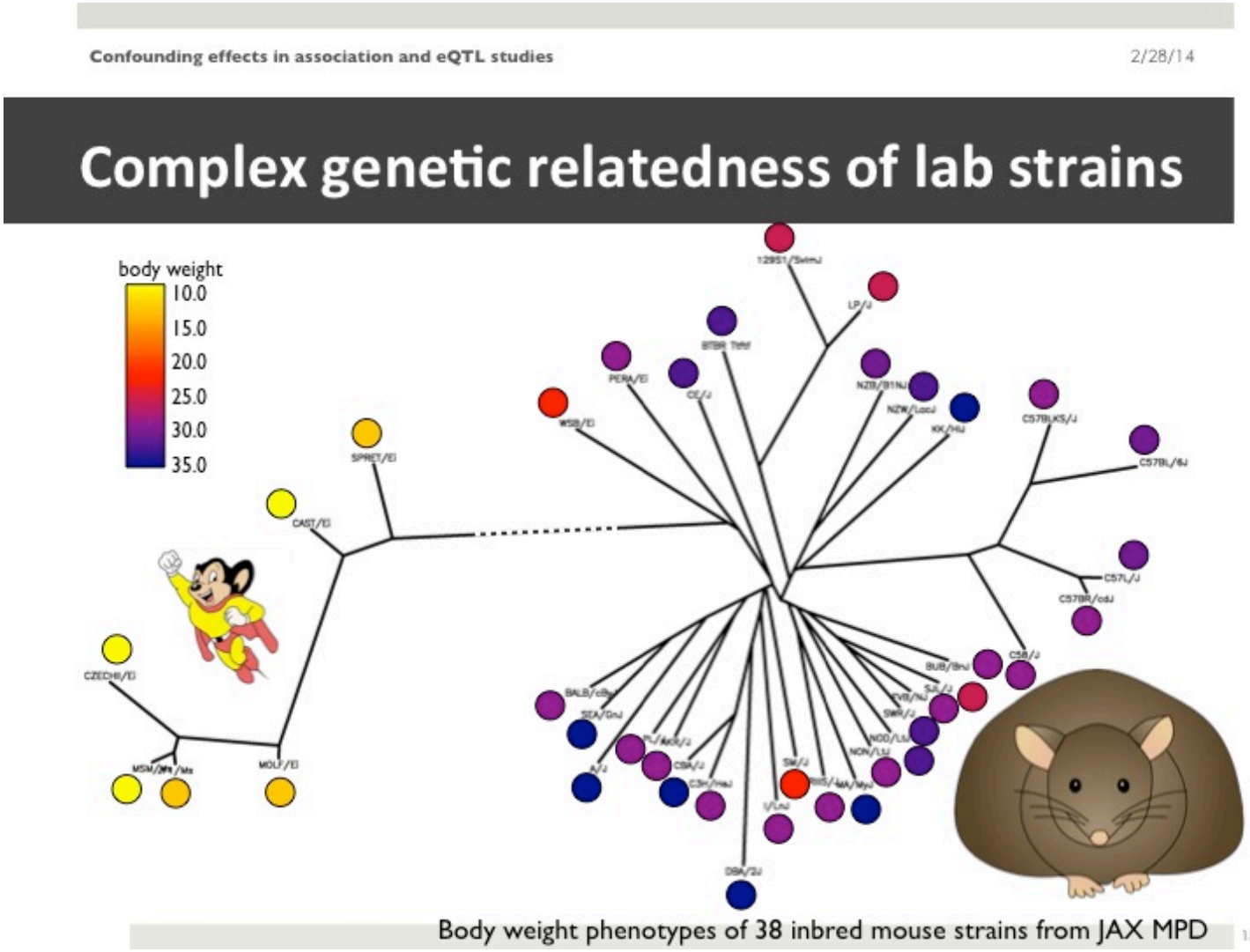
$$K = \frac{1}{M}\sum \frac{(g_{ij} - 2p_i)(g_{ik} - 2p_i)}{2p_i(1-p_i)}$$

- Eigenvalue Decomposition:

$$K = V\lambda V^{-1}$$

- V are the eigenvectors sorted in order of variance explained of K and eigenvalue λ

Random effect models also correct for confounding



Random effect models include a term for noise with a correlation structure

Linear Model

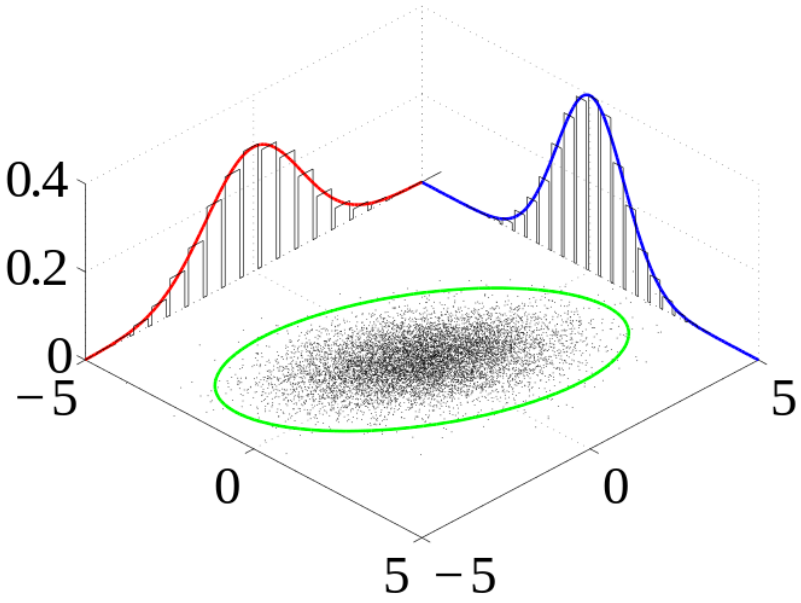
$$y = \sum g_i \beta_i + \varepsilon; \varepsilon \sim N(0,1)$$

Linear Mixed Model

$$y = \sum g_i \beta_i + u + \varepsilon$$

$$\varepsilon \sim N(0,1)$$

$$u \sim MVN(\mathrm{M}, \Sigma)$$



MVN CDF:

$$(2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})},$$

$$\Sigma = K\sigma_g^2 + I\sigma_\varepsilon^2$$

Selected to represent the population!

Linear Mixed Models can also be used to estimate the additive heritability of trait

$$\varphi_j = \sum_{i \in Obs} \beta_i \overline{g}_{ij} + \varepsilon_j$$

$$\beta_i \sim N\left(\mathbf{0}, \mathbf{I} h_g^2\right) \text{ and } \varepsilon_j \sim N\left(\mathbf{0}, \mathbf{I} \sigma_e^2\right)$$

which is equivalent to

$$\varphi_j = u_j + \varepsilon_j$$

$$u_j \sim N\left(\mathbf{0}, K h_g^2\right), \text{ where } K = \frac{1}{N} \overline{g} \overline{g}^T$$

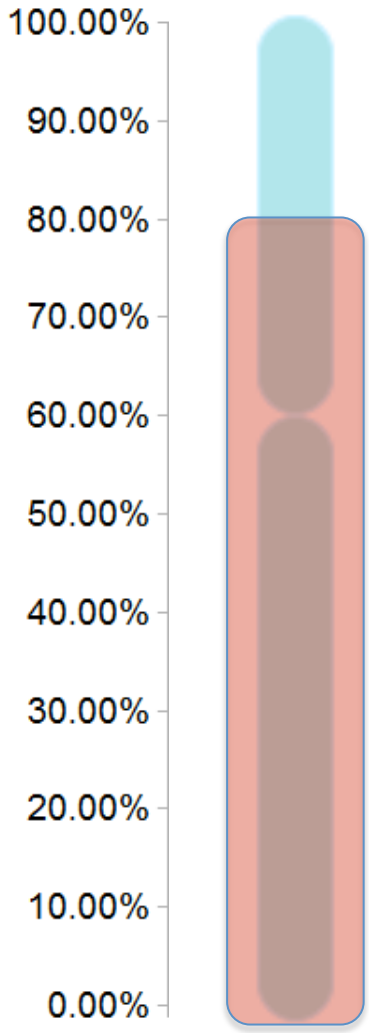
Most textbooks on statistical genetics; Kang (2008) *Genetics*; Kang (2010) *Nat Genet*; Yang (2010) *Nat Genet*; Yang (2011) *Am J Hum Genet*

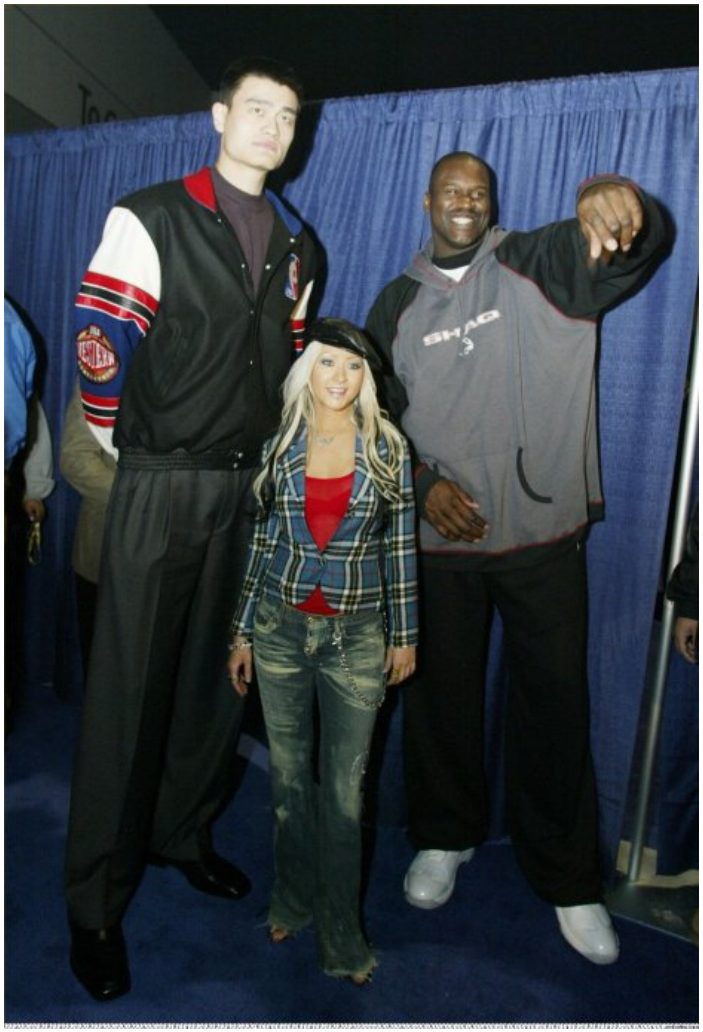
Where is the missing heritability?

- Structural Variants
- Rare Variants
- Weak effects
- Parent of Origin Effects
- Gene-Environment & Gene-Gene Interactions
- Biased heritability estimates

Maher 2008 Nature; Manolio et al. 2009 Nature, Gibson 2012 Nature Reviews Genetics

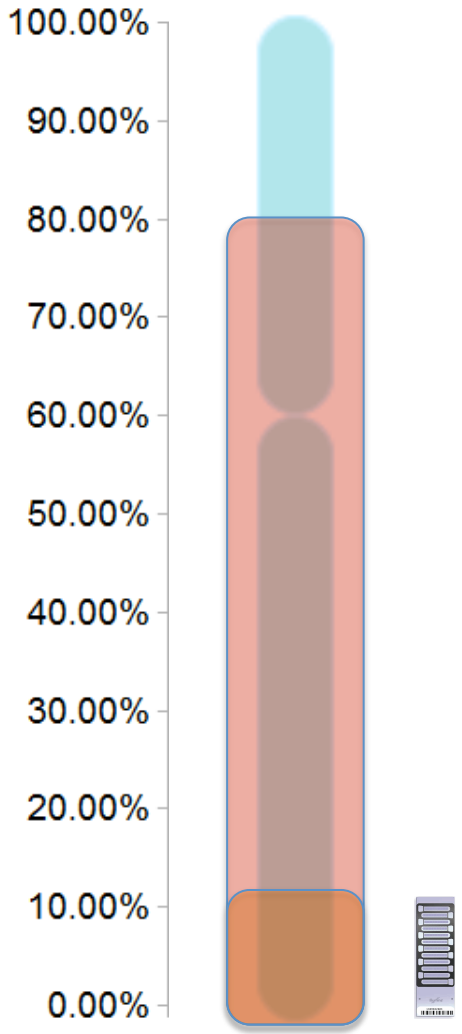
Estimated variation in height driven by genetic differences = 80%

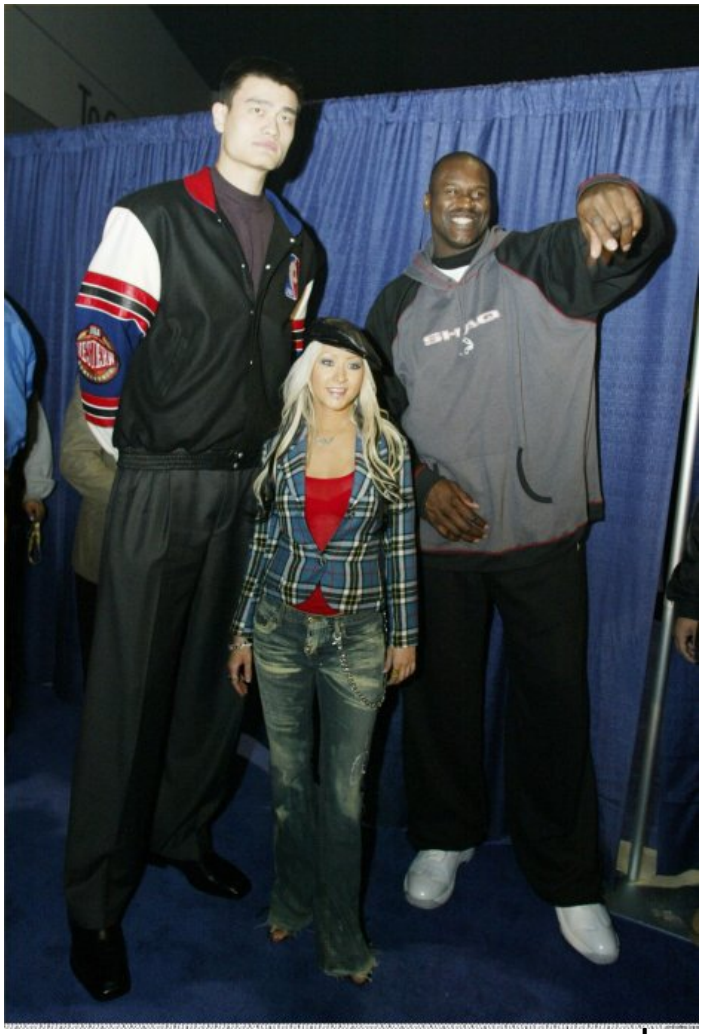




Silventoinen et al, 2003 Twin Research

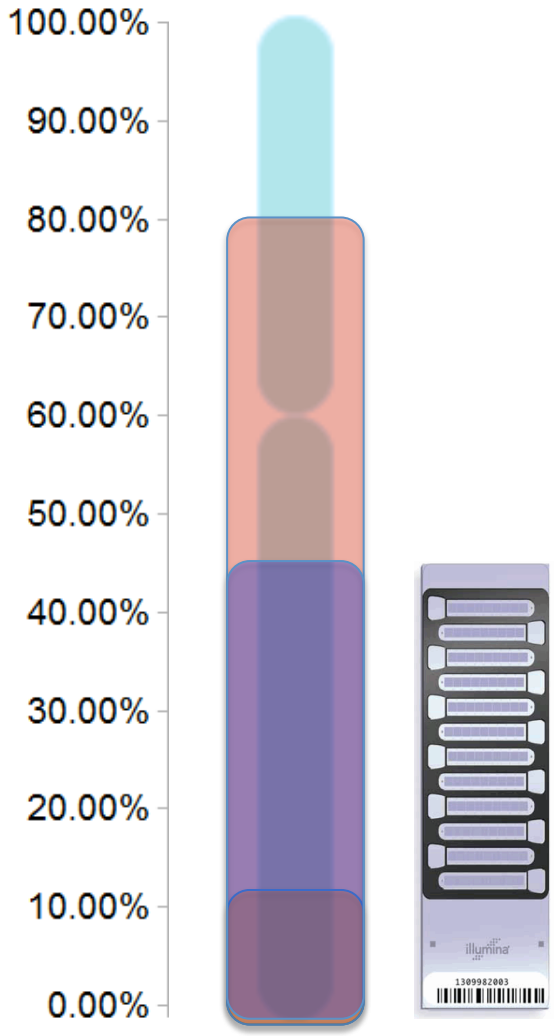
GWAS results explain approximately 10% of variation in height

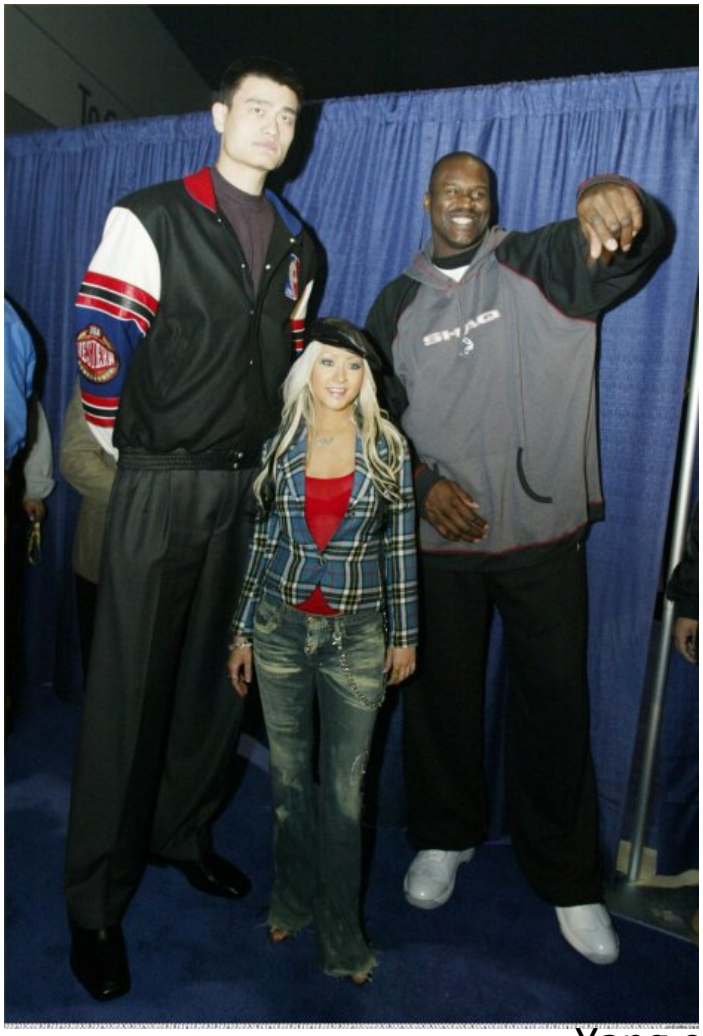




Lango Allen et al, 2010 Nature

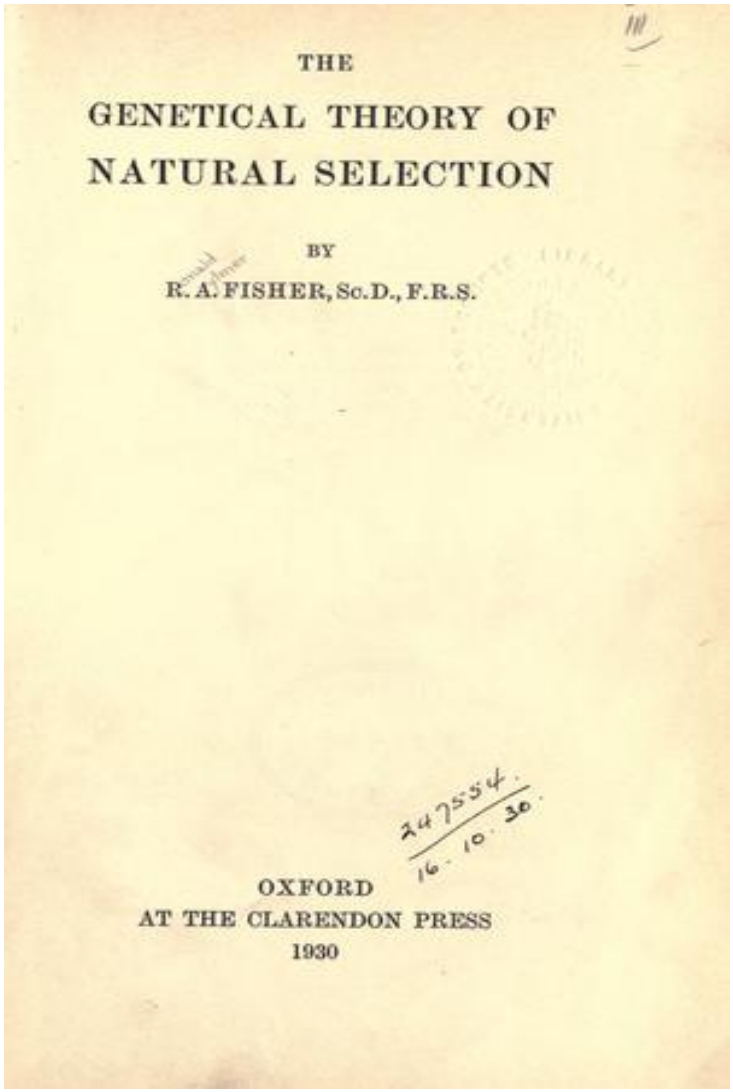
GWAS have the potential to explain 45% of the variation in height





Yang et al, 2011 Nature Genetics

Heritability: Proportion of phenotypic variation driven by genetic variation


THE

GENETICAL THEORY OF

NATURAL SELECTION

BY

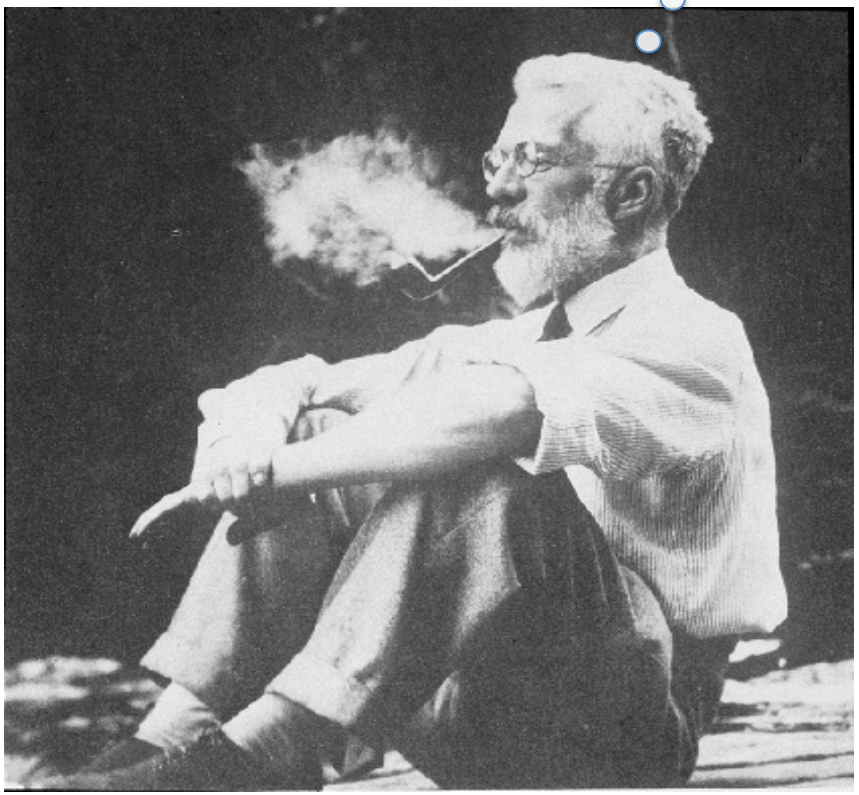
R. A. FISHER, Sc.D., F.R.S.

OXFORD

AT THE CLARENDON PRESS

1930

Phenotype = F(Genes,Environment)



Formal Definition(s) of Heritability

- Phenotype(ϕ) = Genotypes(G) + Environment(ε)

$$\sigma_{\varphi}^2 = \sigma_G^2 + \sigma_{\varepsilon}^2$$

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

$$\text{Narrow Sense}: \; h^2 = \frac{\sigma_A^2}{\sigma_{\varphi}^2} \qquad \text{Broad Sense}: \; H^2 = \frac{\sigma_G^2}{\sigma_{\varphi}^2}$$

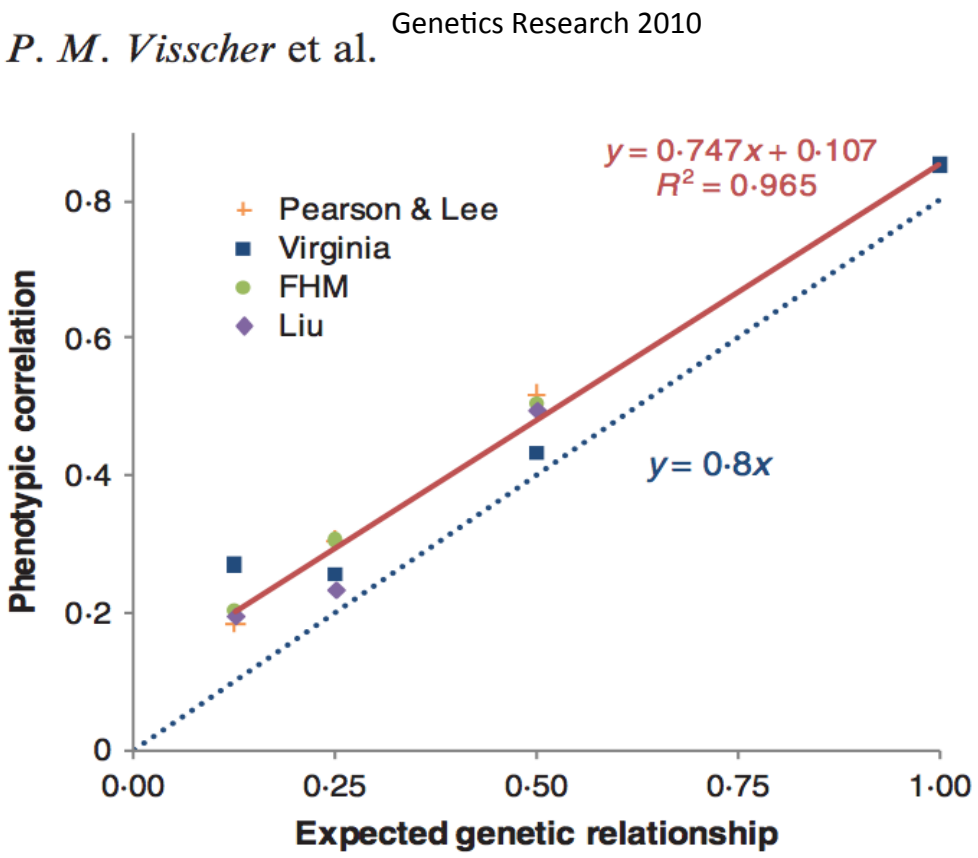
What can be explained by current GWAS analysis methods

What can be explained with the best possible genetic model

How do we estimate heritability?

Solution 1: Use Relatives!

- Measure the phenotypic correlation of twins, siblings, cousins, etc.
- Phenotypic similarity is a function of genetic relationship and heritability.
- By fixing the the relationship class, we can derive the heritability



Visscher et al Nature Reviews Genetics 2008

Heritability Estimation Model

$$\varphi_j : \text{ phenotype} = \sum_i \beta_i \bar{g}_{ij} + \varepsilon_j$$

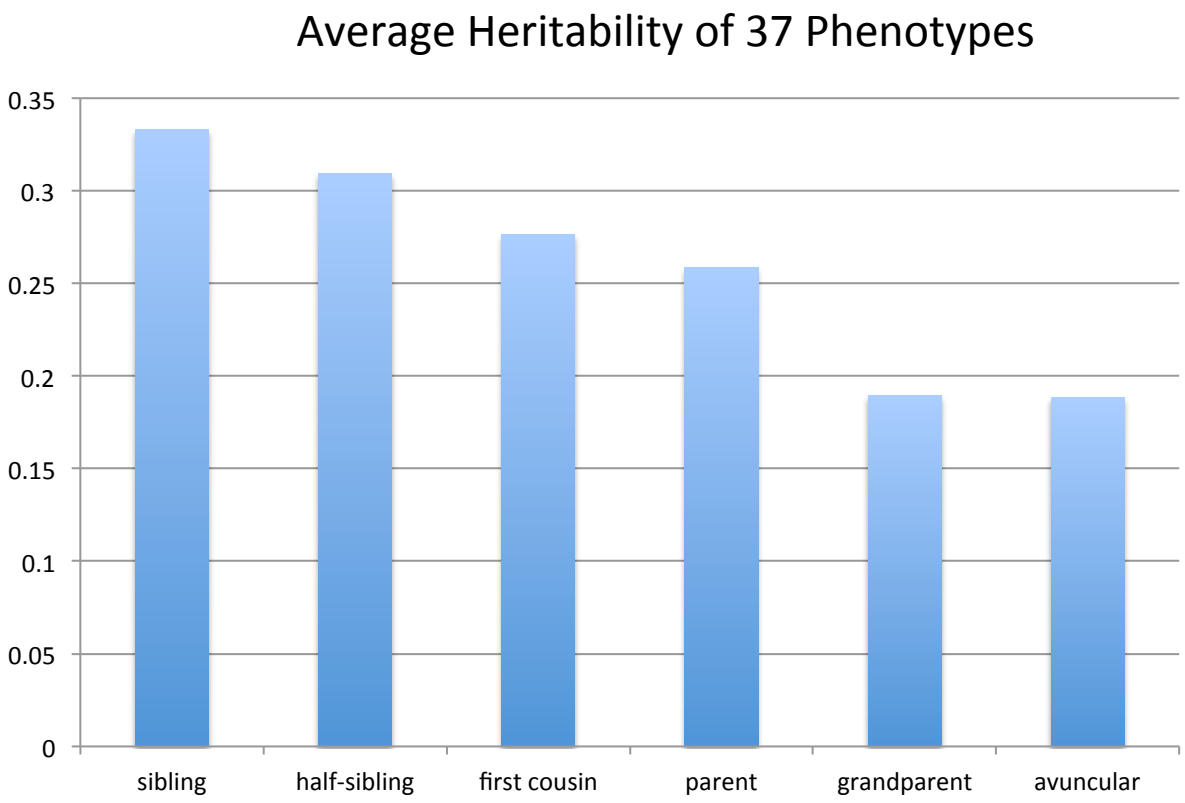
$$\text{cov}(\varphi_j, \varphi_k) = \text{cov}(\sum_i \beta_i \bar{g}_{ij\,i}, \sum_i \beta_i \bar{g}_{ik}) = \frac{h^2}{N} \sum_{i \in C} \frac{\text{cov}(\bar{g}_{ij}, \bar{g}_{ik})}{\text{var}(\bar{g}_i)} = h^2 K[j,k]$$

Unknown

Estimated from data set

Genetic Relationship
E[K] = 0.5 for siblings
E[K] = 0.125 for 1st cousins

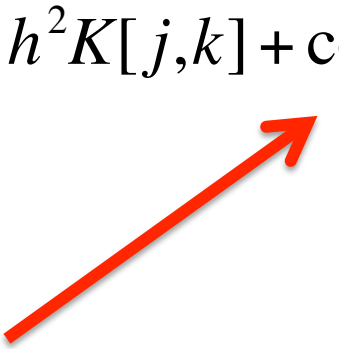
Environmental sharing significantly influences phenotype



Shared environment inflated covariance and heritability estimates

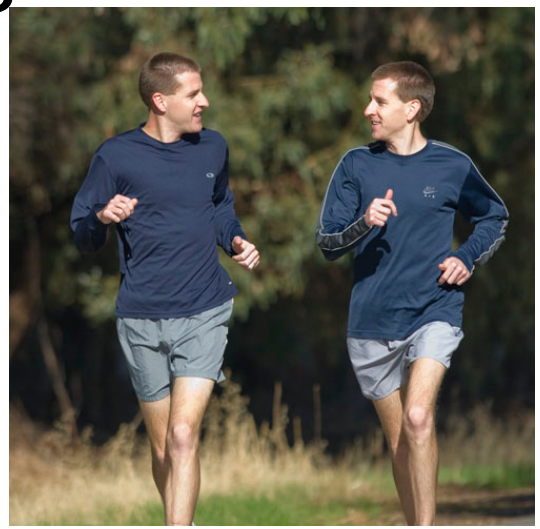
$$\varphi_j : \text{ phenotype} = \sum_i \beta_i \bar{g}_{ij} + \varepsilon_j$$

$$\text{cov}(\varphi_j, \varphi_k) = \text{cov}(\sum_i \beta_i \bar{g}_{ij_i}, \sum_i \beta_i \bar{g}_{ik}) + \text{cov}(\varepsilon_j, \varepsilon_k) = h^2 K[j,k] + \text{cov}(\varepsilon_j, \varepsilon_k)$$



unknown

Problem 1: Relatives share environments as well as genetics

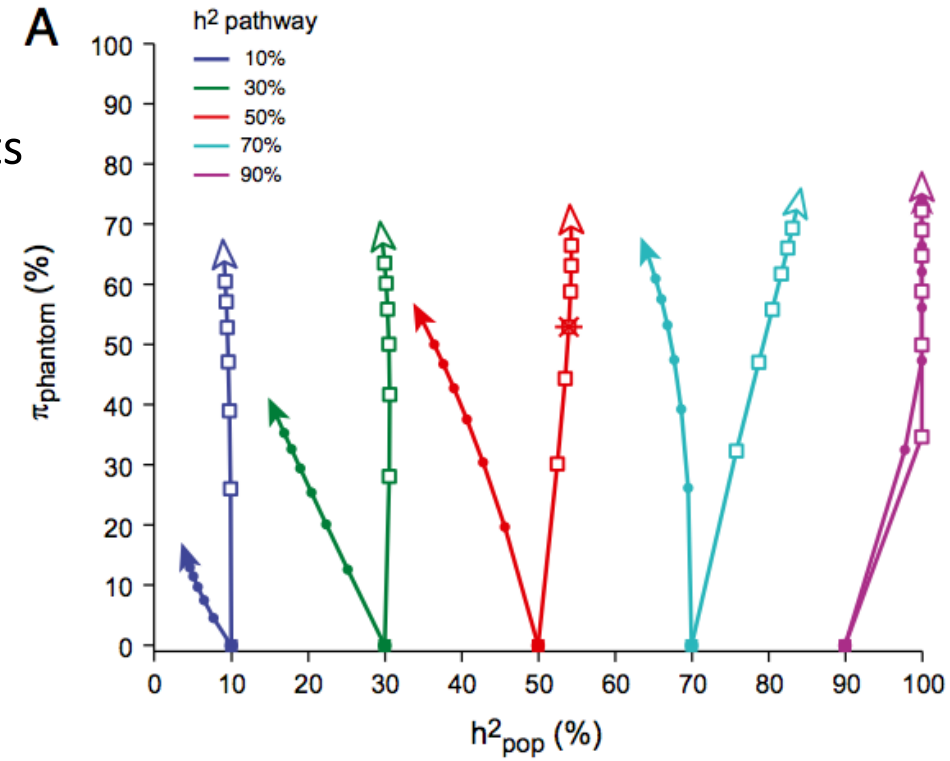


- Phenotypic correlation is a function of shared genetics, shared environment, and heritability.
- Genealogical estimates of shared environmental effects are therefore inflated
- Solution: Compare MZ and DZ twins under the assumption of shared environmental equivalence
  - Poor assumption & confounded by dominance and interactions

Zaitlen et al, 2012a Under Review
Zaitlen et al, 2012b Heritability in the GWAS, Human Heredity
Zaitlen et al 2011 ASHG Platform Talk and Research Trainee Award

Problem 2: Inflated estimates of heritability due to dominance & epistasis

- Interactions are non-linear effects involving multiple SNPs
- Interactions are inherited at different rates than main effects
- $H^2 \neq h^2$
- Narrow-sense heritability estimates even from MZ/DZ twin studies are inflated under interaction models
- Solution: Use unrelated individuals.



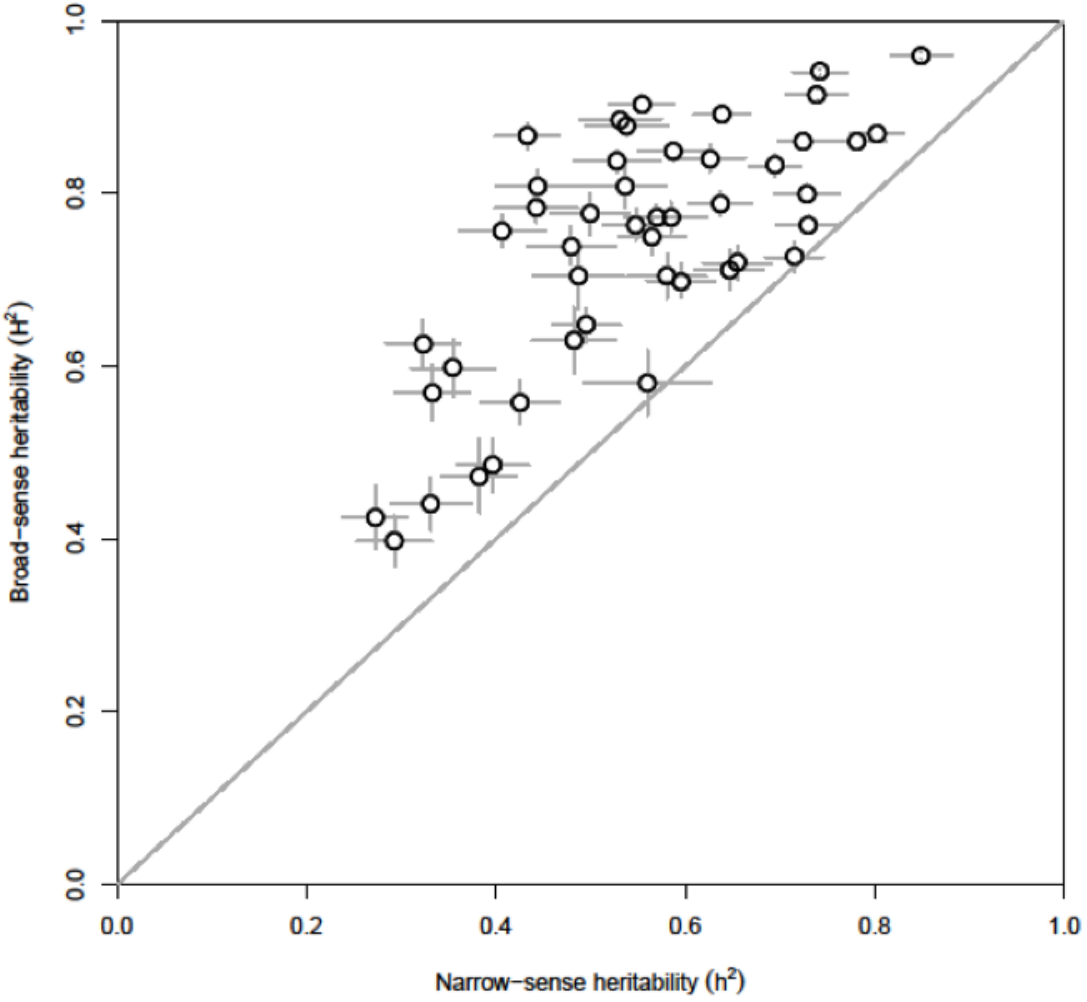
Zuk et al PNAS 2012

Epistasis inflates narrow-sense heritability estimates

$$\varphi_j : \text{phenotype} = \sum_i \beta_i \bar{g}_{ij} + \varepsilon_j + \sum_{i,l} \beta_{il} \bar{g}_{ij} \bar{g}_{lj} + other$$

Sharing of interaction terms is non-linear in relationship.
Pair-wise interaction heritability descends quadraticlly.

Model organisms show many phenotypes with significant epistasis



What does learning broad-sense versus narrow-sense heritability teach us?

Bloom et al arxiv 2012

Hard and Fun Problems in Medical Genetics

- Missing Heritability Problem
- Cross-population heritability problem
- GWAS in admixed populations like African-Americans and Latinos
- Phenotypic prediction
- Parent of Origin effects and Imprinting
- Environmental, genomic, bacterial, and other external data sources