# Evolution, maintenance and allelic architecture of complex traits

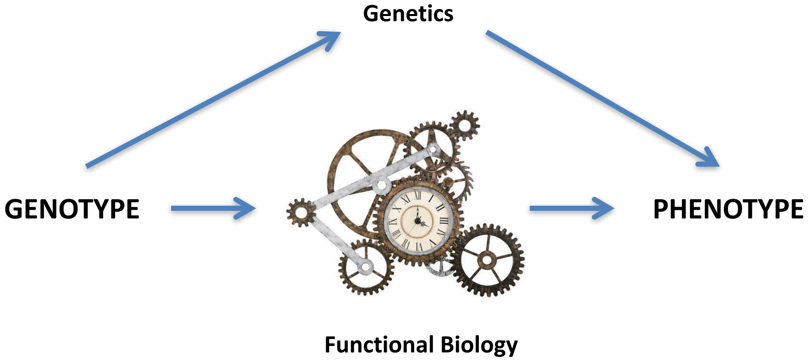
Shamil Sunyaev

**Department of Biomedical Informatics**
Harvard Medical School

**Division of Genetics**
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

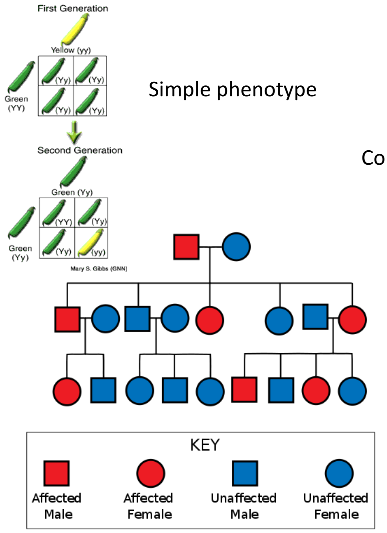
## Why are we doing genetics?



# The role of statistics

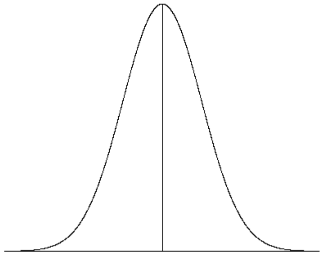
- Genetics for statistics is what physics is for mathematics
- Genetics is a leading motivation for development of new basic statistics
- Statistics is the main formal instrument (although not the only one)

## Simple and Complex Phenotypes



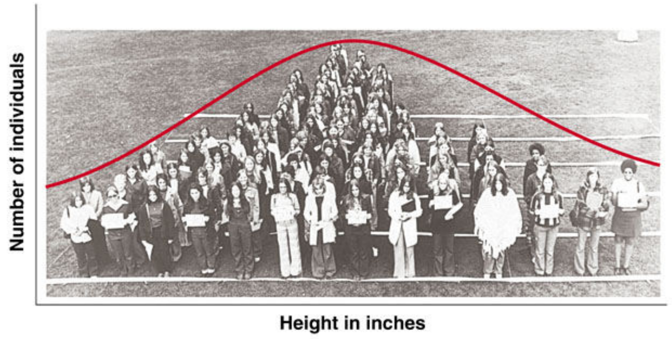
Simple phenotype

Complex phenotype

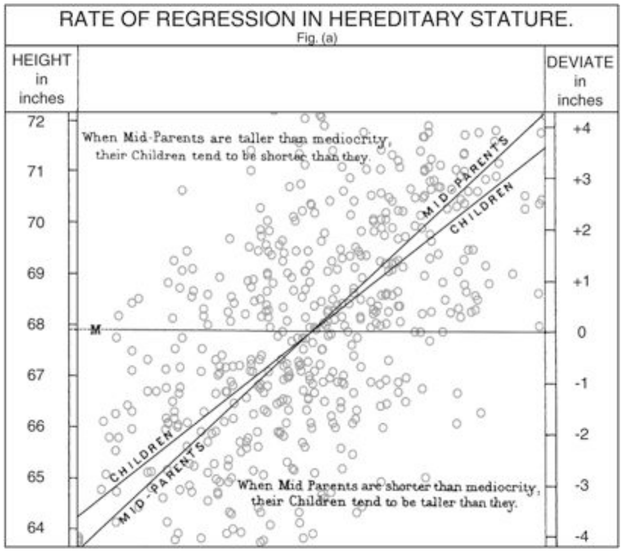


# Complex traits are heritable but not in Mendelian fashion

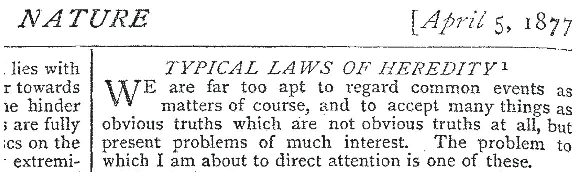
Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6



# Complex traits are heritable but not in Mendelian fashion



# Infinitesimal model

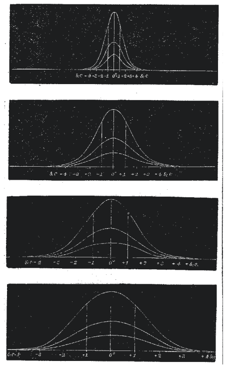


*NATURE* [*April* 5, 1877

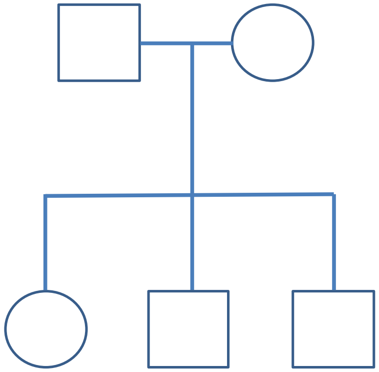
. lies with
r towards
ıe hinder
; are fully
;cs on the
· extremi-

*TYPICAL LAWS OF HEREDITY*[1]

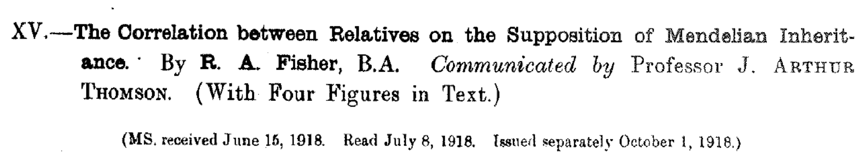
WE are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest. The problem to which I am about to direct attention is one of these.



# Infinitesimal model: multivariate normal distribution in pedigrees



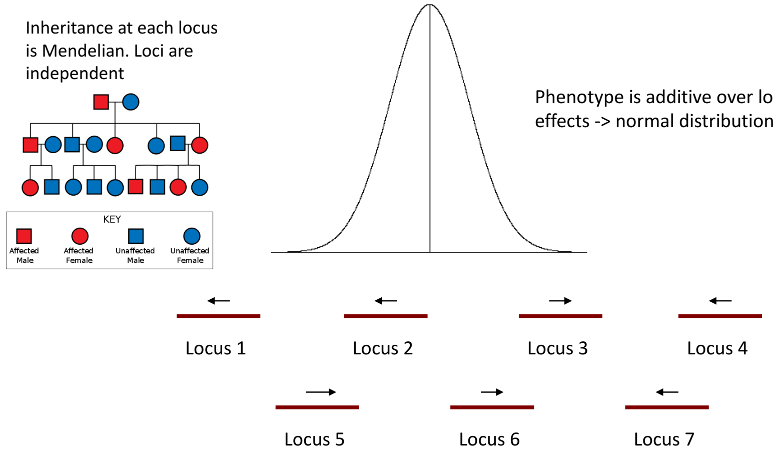
The pedigree defines the covariance matrix



XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. By R. A. Fisher, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

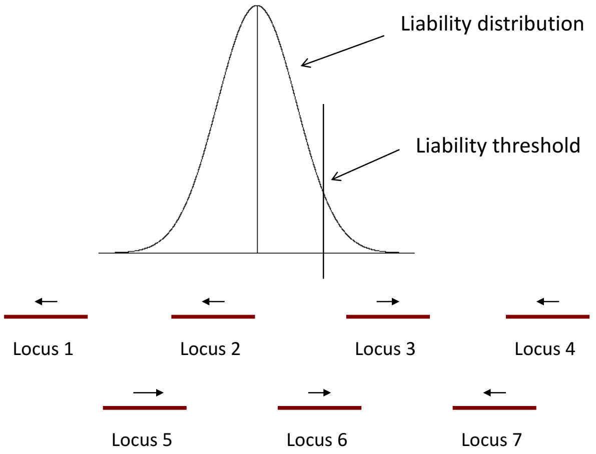
## Quantitative Trait Loci (QTLs)



Inheritance at each locus is Mendelian. Loci are independent

Phenotype is additive over locus effects -> normal distribution

## Dichotomous complex traits such as disease

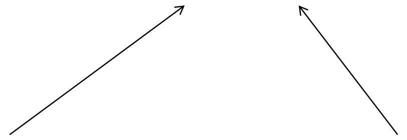


Liability distribution

Liability threshold

## Population variation is fully described by variance

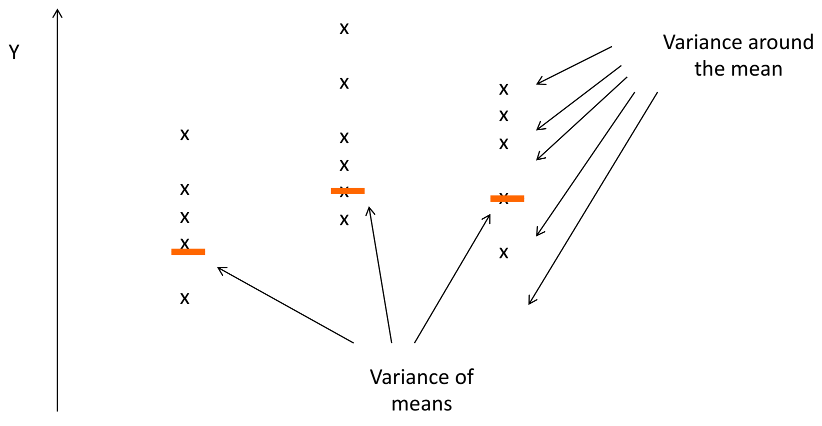
$$V = V_G + V_E$$



Genetic contribution

Everything else

# Variance decomposition

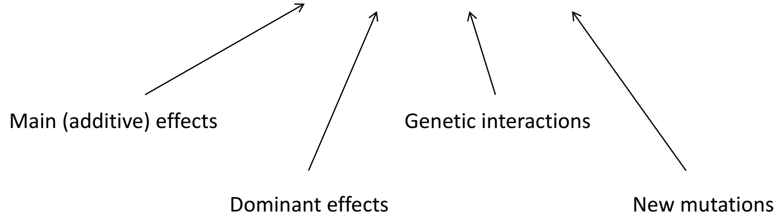


Variance around the mean

Variance of means

# Components of genetic variance

$$V_G = V_A + V_D + V_I + V_M$$

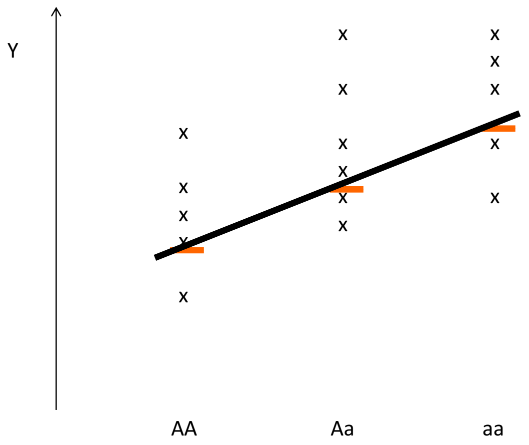

Main (additive) effects

Dominant effects

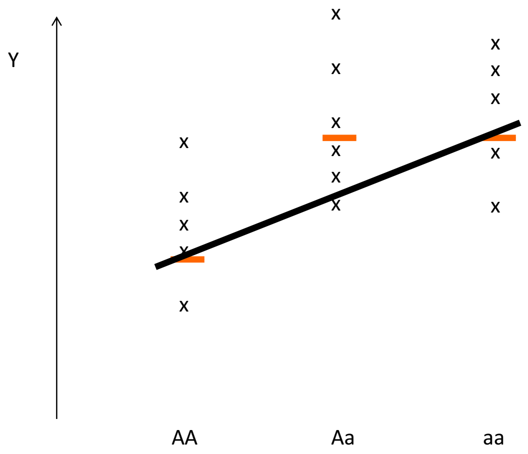
Genetic interactions

New mutations

# Regression



# Regression



# Additive variance

Additive variance $V_A$ is variance explained by the model

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

$$V_A = 2\sum_i \beta_i^2 x_i \left(1 - x_i\right)$$

# Variance components due to dominance and epistasis

Dominance variance $V_D$ is variance explained by the residuals of the model additive over loci

Epistatic variance $V_I$ is genetic variance that is not captured by the model additive over loci (presumably due to interactions)

Additive by additive pairwise epistasis

$$Y_j = \sum_i \beta_i X_{ij} + \sum_{lk} \beta_{lk} X_{lj} X_{kj} + \varepsilon$$

# Other variance components

Epistasis can be additive by dominant and dominant by dominant

Epistasis can be due to higher order interactions

Mutational variance $V_M$ – additional variance due to *de novo* mutations

# Heritability

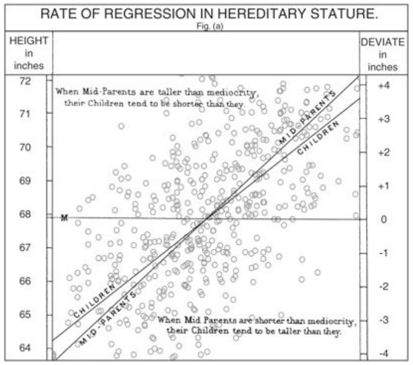
Broad sense

$$H^2 = \frac{V_G}{V}$$

Narrow sense

$$h^2 = \frac{V_A}{V}$$

# Estimating heritability

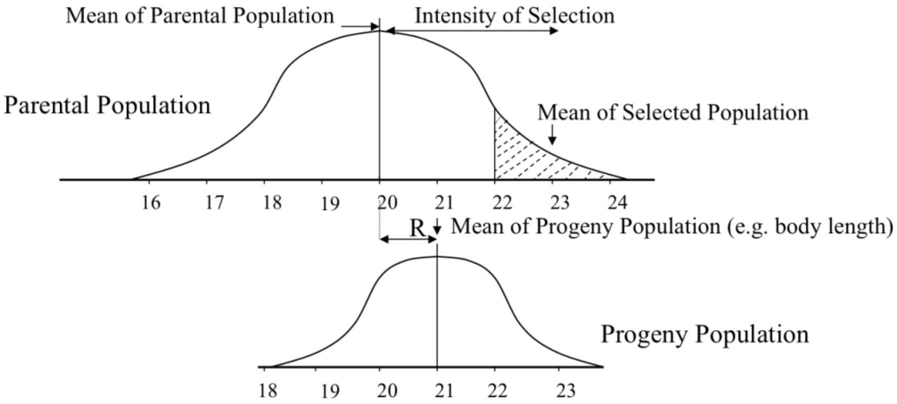


$$Cov(MP,O) = \frac{1}{2}V_A + \frac{1}{4}V_I$$

Narrow sense heritability

$$h^2 = \frac{V_A}{V} \approx \frac{Cov(MP,O)}{V(MP)}$$

# Breeder's equation

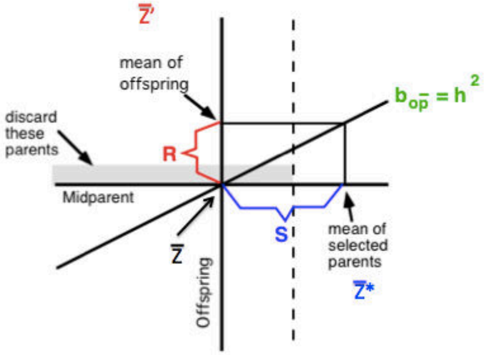

Mean of Parental Population
Intensity of Selection
Parental Population
Mean of Selected Population
16
17
18
19
20
21
22
23
24
R
Mean of Progeny Population (e.g. body length)
Progeny Population
18
19
20
21
22
23

$$R = h^2 S$$

# Breeder's equation

Response to Selection = heritability * Selection Differential

$$R \quad = \quad h^2 \quad S$$



$$h^2 = \frac{V_A}{V_P}$$

# With genotypic information in hand

Regress phenotype on genotype

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

Additive variance

$$V_A = 2\sum_i \beta_i^2 x_i \left(1 - x_i\right)$$

Narrow sense heritability

$$h^2 = \frac{V_A}{V}$$

# In the Ideal World

Regress phenotype on genotype

$$Y_j = \sum_i \beta_i X_{ij} + \varepsilon$$

Identify significant and reproducible associations.
Estimate effect sizes. Estimate additive variance.

$$\hat{V}_A = 2 \sum_i^{known} \hat{\beta}_i^2 \, x_i \left(1 - x_i\right)$$

Reality: missing heritability

$$\hat{h}^2 = \frac{\hat{V}_A}{V} << \frac{Cov(MP,O)}{V(MP)}$$

# Current GWAS explain a minor fraction of heritability

The case of the missing heritability

Height – 10%, Blood lipids – 12%

# Likely reasons for missing heritability

1. *Common variants of weak effect*
2. *Rare variants of larger effect*
3. *Epistatic interactions*

$$Cov(MP,O) = \frac{1}{2}V \boxed{+\frac{1}{4}V_I}$$

# Questions about allelic architecture

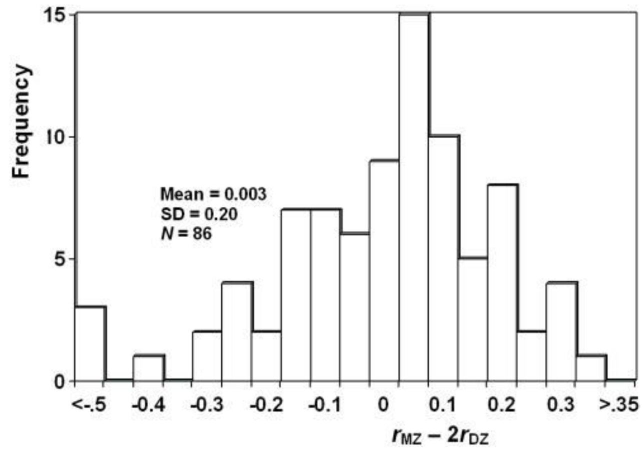
- How many loci are involved?
- Is variation underlying the trait rare or common?
- What is the distribution of effect sizes of variants involved in the trait?
- What is the role of epistasis and dominance?

*GxG interactions*

## Why is epistasic variance commonly disregarded?

- In human genetics, epistatic interactions between common variants have not been observed.
- In a model with two (or several) loci, contribution of epistatic variance is relatively small.
- Long term response to selection in model organisms seems to contradict the importance of epistasis.

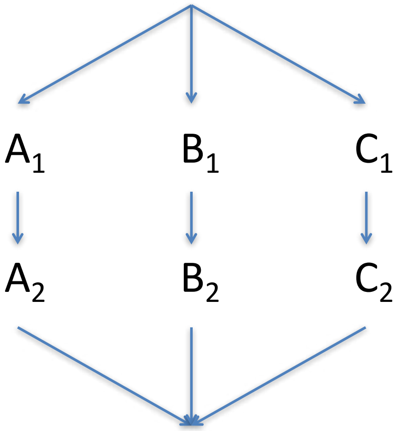
## Any evidence for or against epistasis?



## Why is epistasic variance might be of importance?

- A non-linear model involving many loci would generate a large epistatic variance.
- Interactions would be statistically undetectable.
- The model would not generate significant deviations from the observations.
- As an example, we may consider a model with multiple pathways involved.

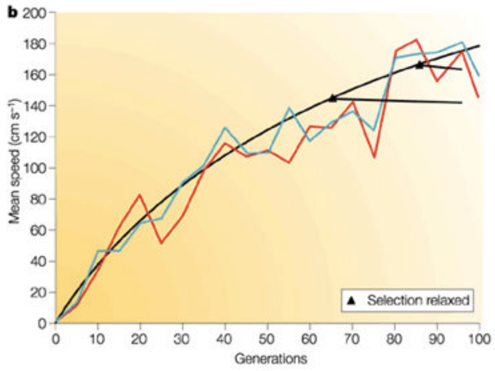
## Multiple pathway model



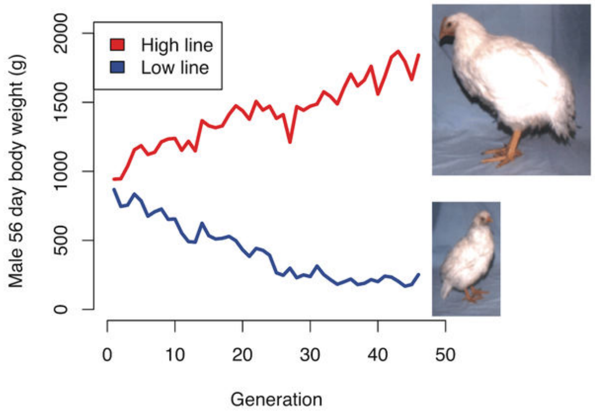
$A_1$ $B_1$ $C_1$

$A_2$ $B_2$ $C_2$

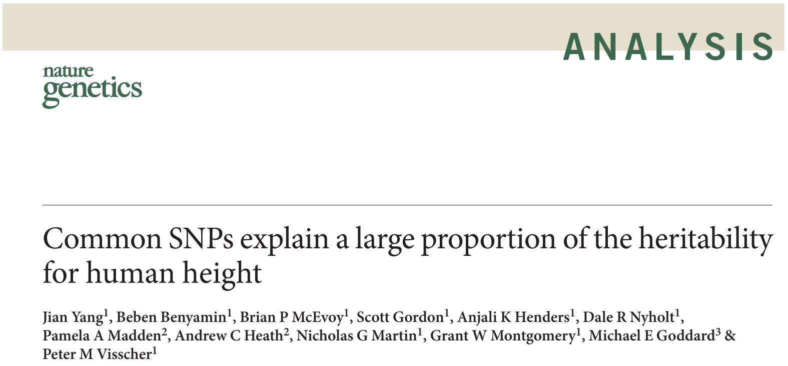
## Evidence in favor of the highly polygenic model



## Evidence in favor of the highly polygenic model



## Evidence in favor of the highly polygenic model



ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

| | AA | Aa | aa |
|---|---|---|---|
| Genotypes | 0 | 1 | 2 |
| Normalized genotypes | $\frac{0-E(X)}{\sqrt{Var(X)}}$ | $\frac{1-E(X)}{\sqrt{Var(X)}}$ | $\frac{2-E(X)}{\sqrt{Var(X)}}$ |
| Normalized genotypes | $\frac{-2q}{\sqrt{2pq}}$ | $\frac{p-q}{\sqrt{2pq}}$ | $\frac{2p}{\sqrt{2pq}}$ |

If SNP1 is causal and SNP2 is not, the apparent association of SNP2 is:

$$\hat{\beta}_2 = \beta_1 \cdot r_{12}$$

In non-normalized genotypes

$$\hat{\beta}_2 = \beta_1 \cdot r_{12} \cdot \sqrt{\frac{Var(X_1)}{Var(X_2)}}$$

$X_{ij}$ – Normalized genotype of individual *i* at SNP *j*

In the matrix form:

$$\overline{y} = X\overline{\beta} + \varepsilon$$

Two important matrices:

$$LD = \frac{1}{M} X^T X$$

$$GRM = \frac{1}{N} XX^T$$

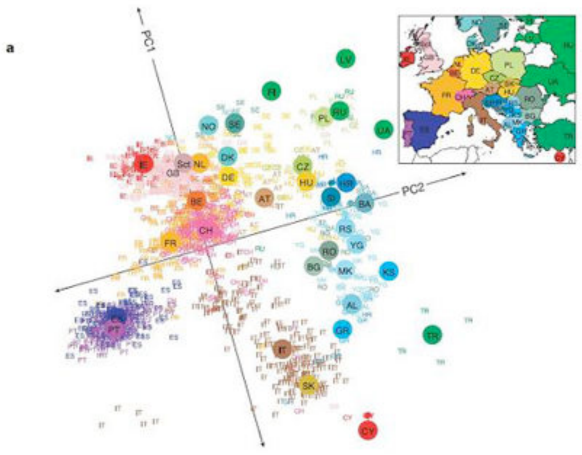
# Principle component analysis (PCA)

$$GRM = \frac{1}{N} XX^T$$

Principle component are eigenvectors

First principle component corresponds to the largest eigenvalue

# Europe



## Linear Mixed Models (LMM)

- We can model effects of individual variants as random effects distributed as *N(0,$\sigma^2$)*.
- Random effect model is a model with error terms drawn from a multivariate normal distribution.
- In the infinitesimal model, co-variance matrix can be approximated using IBS (not IBD).

## Linear Mixed Model (LMM)

Our model

$$Y_i = \sum_j \beta_j X_{ij} + \varepsilon$$

We have to fit markers individually

$$Y_i = \beta_1 X_1 + \sum_{j=2} \beta_j X_{ij} + \varepsilon \sim \beta_1 X_1 + \varepsilon'$$

For each SNP we can fit the model

$$Y_i = \beta X_i + u_i + \varepsilon$$

$$\varepsilon \sim N\left(0, I\sigma^2\right) \qquad u \sim MVN\left(0, GRM\right)$$

## Remember from the Galton plot

Parent and offspring share 50% of DNA (IBD)

$$Cov(P,O) = \frac{1}{2} V_A$$

More generally, if fraction of the genome IBD is ***r***

$$Cov(A,B) = \frac{1}{r} V_A$$

## If we assume that genetic effects are random

We assume that all SNPs have effects on the trait drawn from a normal distribution

$$Y_i = \mu_i + u_i + \varepsilon$$

$$Cov(u_i, u_k) = \frac{1}{N} \sigma^2 \sum_j X_{ij} X_{ik}$$

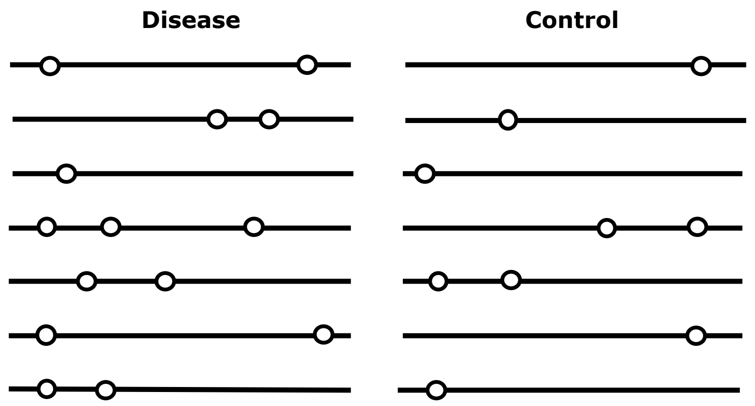
$$u \sim MVN\left(0, \sigma^2 GRM\right)$$

# Challenges of the polygenic model

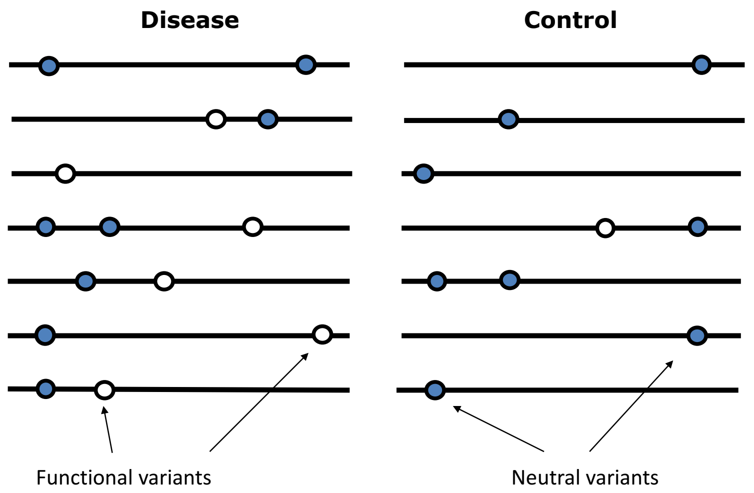
1) Need for a very large target size
2) Natural selection is expected to rapidly eliminate variants and reduce allele frequency of remaining variants
3) Variants must be either very rare or of very small effect sizes

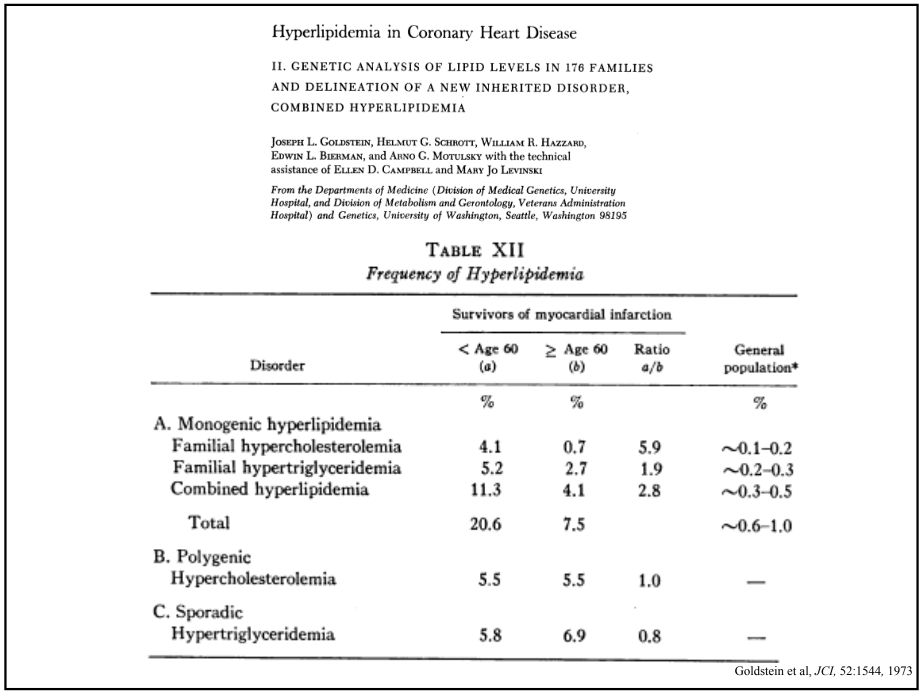
*Rare variants*

## This is a direct association!



## This is a direct association!





Hyperlipidemia in Coronary Heart Disease

II. GENETIC ANALYSIS OF LIPID LEVELS IN 176 FAMILIES AND DELINEATION OF A NEW INHERITED DISORDER, COMBINED HYPERLIPIDEMIA

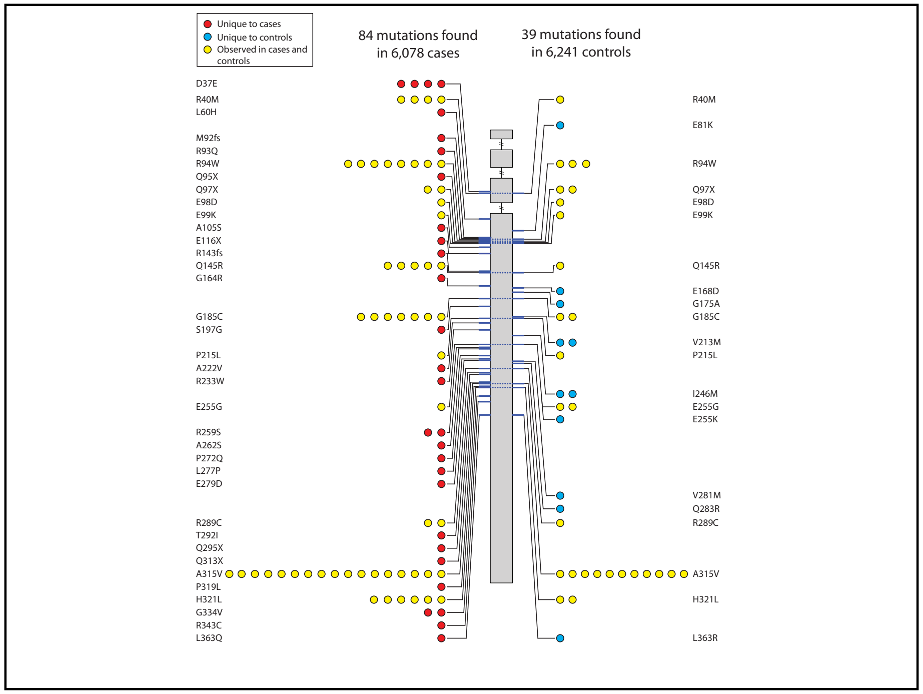
JOSEPH L. GOLDSTEIN, HELMUT G. SCHROTT, WILLIAM R. HAZZARD, EDWIN L. BIERMAN, and ARNO G. MOTULSKY with the technical assistance of ELLEN D. CAMPBELL and MARY JO LEVINSKI

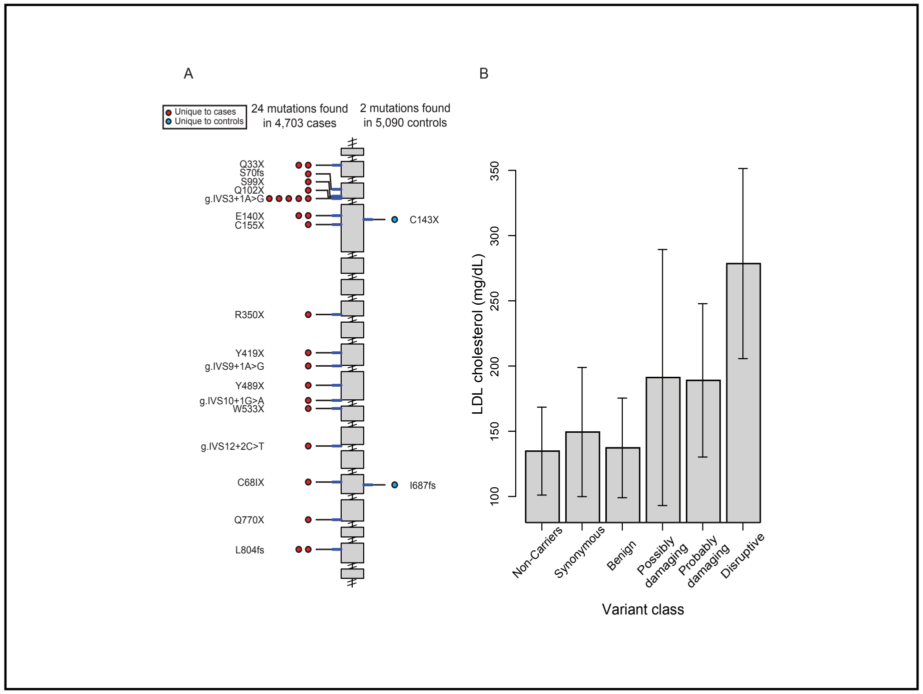
*From the Departments of Medicine (Division of Medical Genetics, University Hospital, and Division of Metabolism and Gerontology, Veterans Administration Hospital) and Genetics, University of Washington, Seattle, Washington 98195*

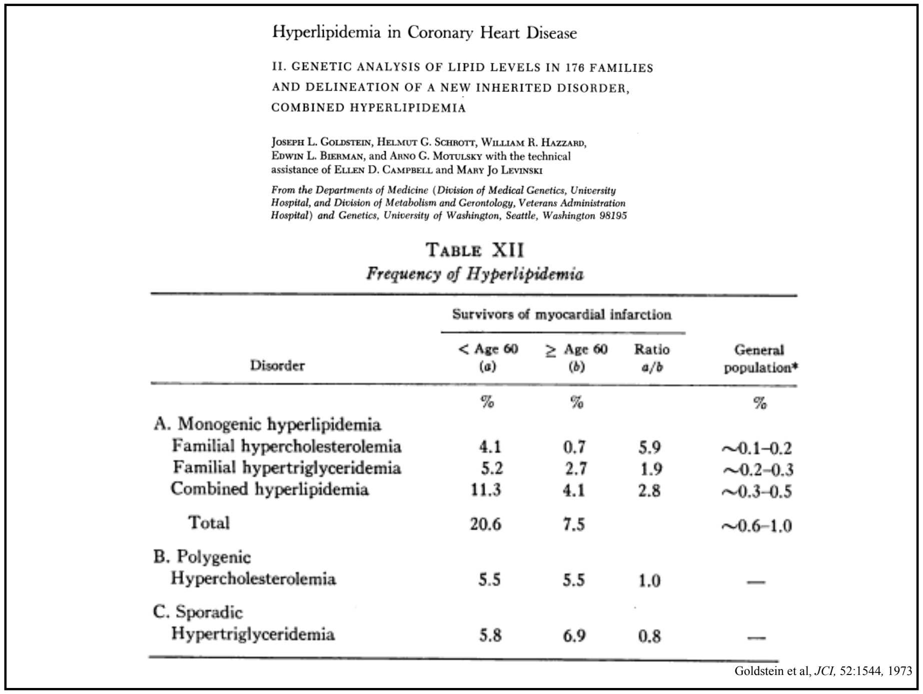
TABLE XII
*Frequency of Hyperlipidemia*

| Disorder | Survivors of myocardial infarction | | | General population* |
|---|---|---|---|---|
| | < Age 60 (*a*) | ≥ Age 60 (*b*) | Ratio *a/b* | |
| | % | % | | % |
| A. Monogenic hyperlipidemia | | | | |
| Familial hypercholesterolemia | 4.1 | 0.7 | 5.9 | ~0.1–0.2 |
| Familial hypertriglyceridemia | 5.2 | 2.7 | 1.9 | ~0.2–0.3 |
| Combined hyperlipidemia | 11.3 | 4.1 | 2.8 | ~0.3–0.5 |
| Total | 20.6 | 7.5 | | ~0.6–1.0 |
| B. Polygenic Hypercholesterolemia | 5.5 | 5.5 | 1.0 | — |
| C. Sporadic Hypertriglyceridemia | 5.8 | 6.9 | 0.8 | — |

Goldstein et al, *JCI,* 52:1544, 1973






Hyperlipidemia in Coronary Heart Disease

II. GENETIC ANALYSIS OF LIPID LEVELS IN 176 FAMILIES AND DELINEATION OF A NEW INHERITED DISORDER, COMBINED HYPERLIPIDEMIA

JOSEPH L. GOLDSTEIN, HELMUT G. SCHROTT, WILLIAM R. HAZZARD, EDWIN L. BIERMAN, and ARNO G. MOTULSKY with the technical assistance of ELLEN D. CAMPBELL and MARY JO LEVINSKI

*From the Departments of Medicine (Division of Medical Genetics, University Hospital, and Division of Metabolism and Gerontology, Veterans Administration Hospital) and Genetics, University of Washington, Seattle, Washington 98195*

TABLE XII
*Frequency of Hyperlipidemia*

| Disorder | Survivors of myocardial infarction | | | General population* |
|---|---|---|---|---|
| | < Age 60 (*a*) | ≥ Age 60 (*b*) | Ratio *a/b* | |
| | % | % | | % |
| A. Monogenic hyperlipidemia | | | | |
| Familial hypercholesterolemia | 4.1 | 0.7 | 5.9 | ~0.1–0.2 |
| Familial hypertriglyceridemia | 5.2 | 2.7 | 1.9 | ~0.2–0.3 |
| Combined hyperlipidemia | 11.3 | 4.1 | 2.8 | ~0.3–0.5 |
| Total | 20.6 | 7.5 | | ~0.6–1.0 |
| B. Polygenic Hypercholesterolemia | 5.5 | 5.5 | 1.0 | — |
| C. Sporadic Hypertriglyceridemia | 5.8 | 6.9 | 0.8 | — |

Goldstein et al, *JCI,* 52:1544, 1973