week 8

genotype to phenotype, GWAS, linear regression

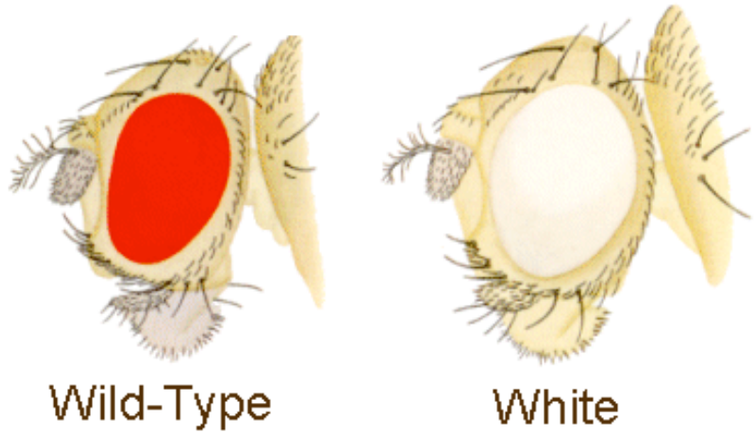
Quantitative Genetics

?

A living histogram. These students and faculty at the University of Connecticut have sorted themselves into columns by height. (Peter Morenus, University of Connecticut)

Linking Genotypic variation to Phenotypic variation

What kind of phenotypes?

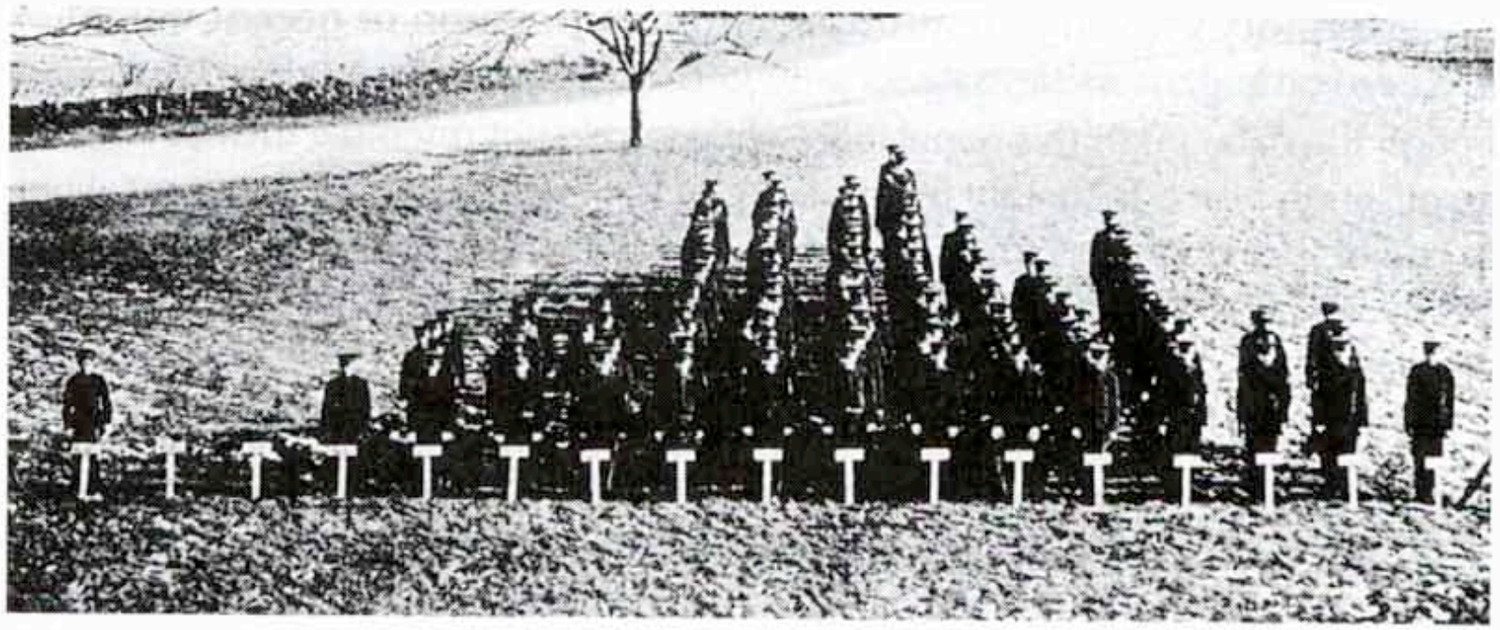


Every kind:

- Mendelian (single locus) traits easier
- Continuous traits harder -- focus of Quantitative Genetics

Body size distribution in a marching band.

(a)



(b)

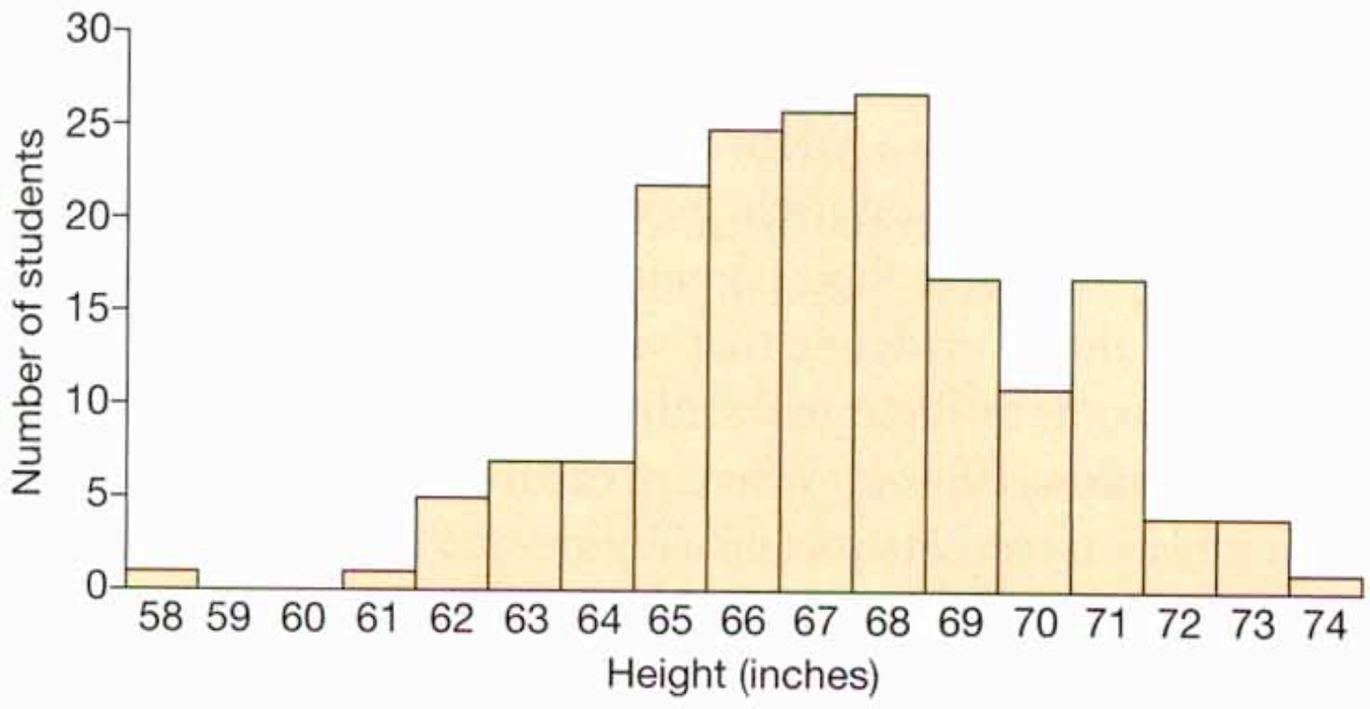


Figure 7.19 Normally distributed variation in a trait (a) A photograph, published in the Journal of Heredity in 1914 by Albert Blakeslee, of a group of students at Connecticut Agricultural College sorted by height. The arrangement of the students forms a living histogram. (Compare Blakeslee's photo to the one on page 197, taken at the same school in 1996.) Reproduced by permission of Oxford University Press (b) A graphical histogram representing the distribution of heights among the students shown in (a).

Freeman and Herron 2001.

Body size distribution in a marching band.

A living histogram. These students and faculty at the University of Connecticut have sorted themselves into columns by height. (Peter Morenus, University of Connecticut)

Freeman and Herron 2001.

How do we measure variation?

Continuous Variation

Mean: $\frac{1}{n}\sum X_i$

Variance: $\frac{1}{n-1}\sum (\bar{X} - X_i)^2$

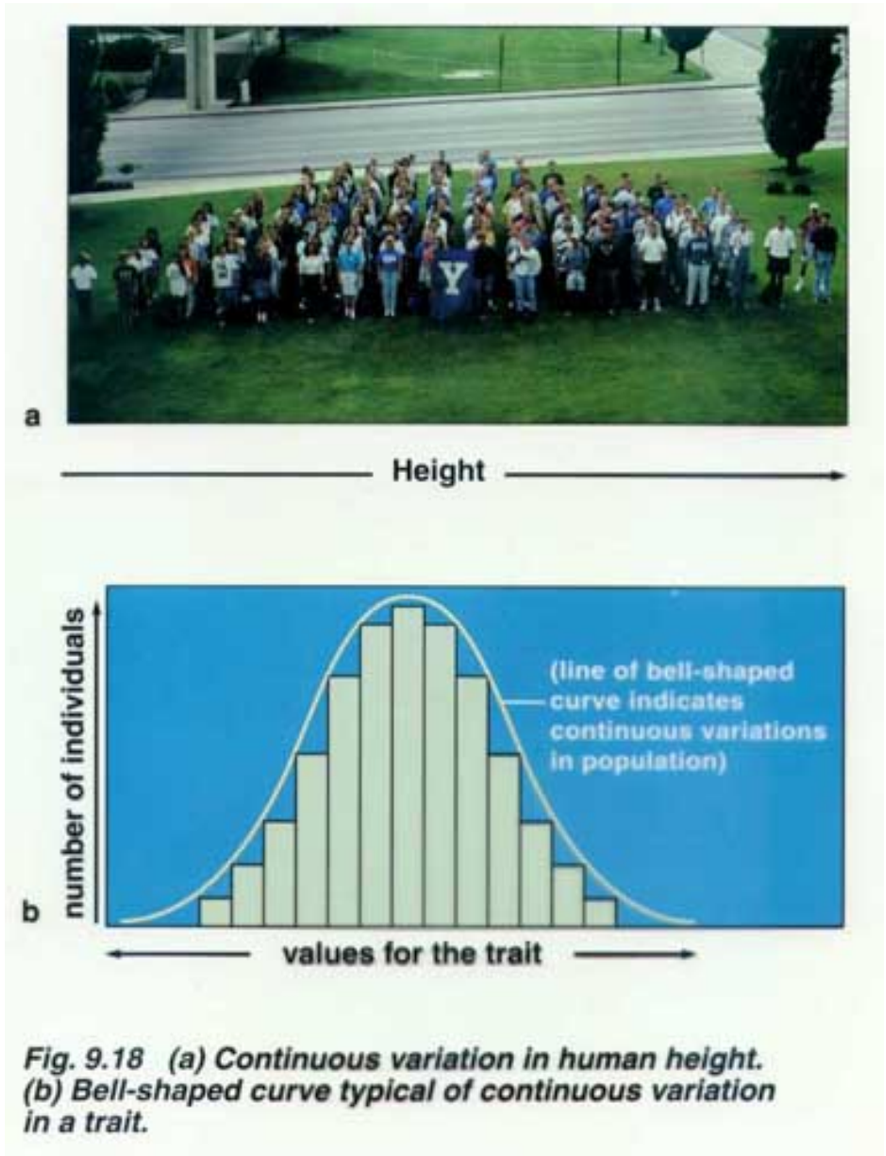
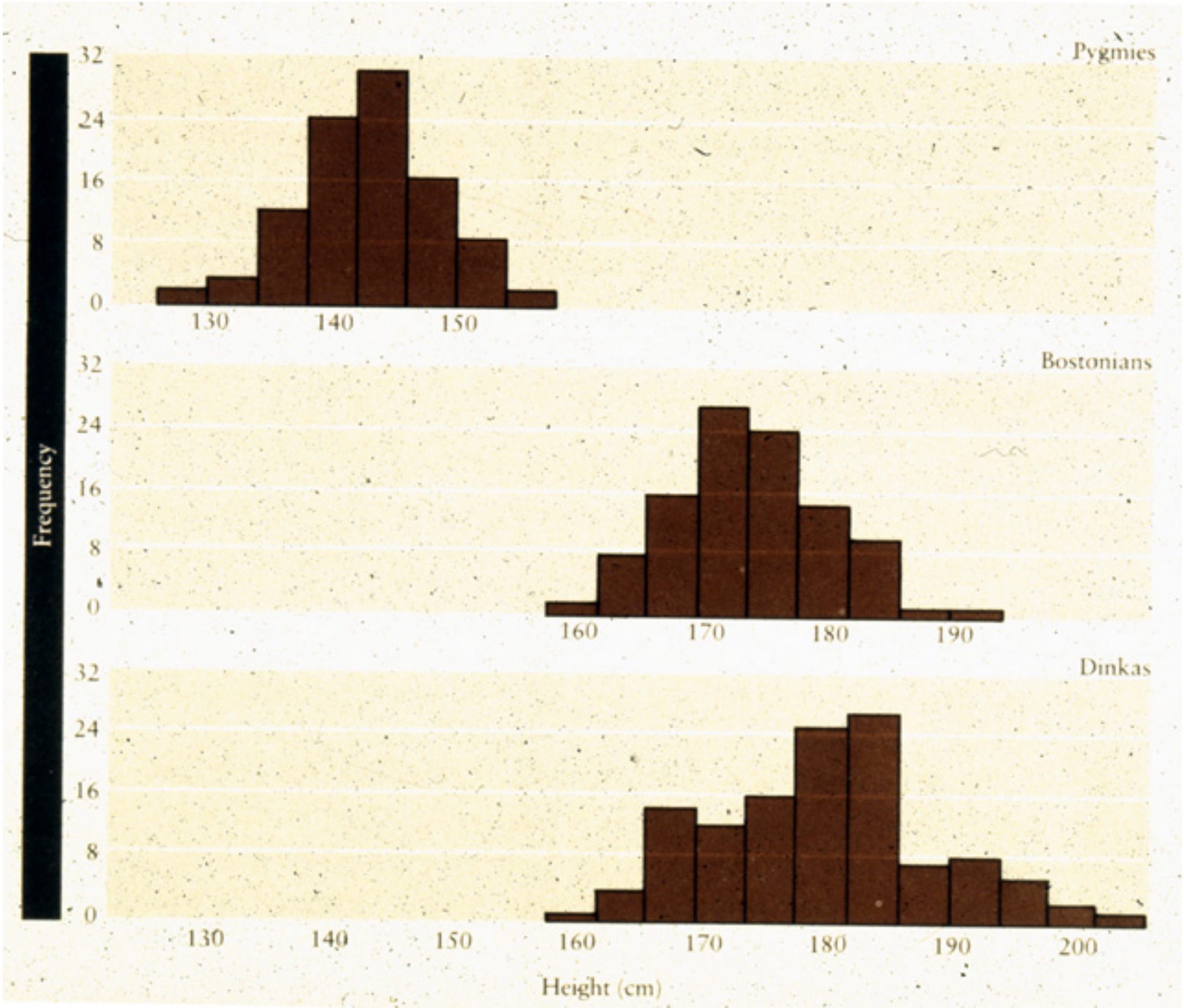


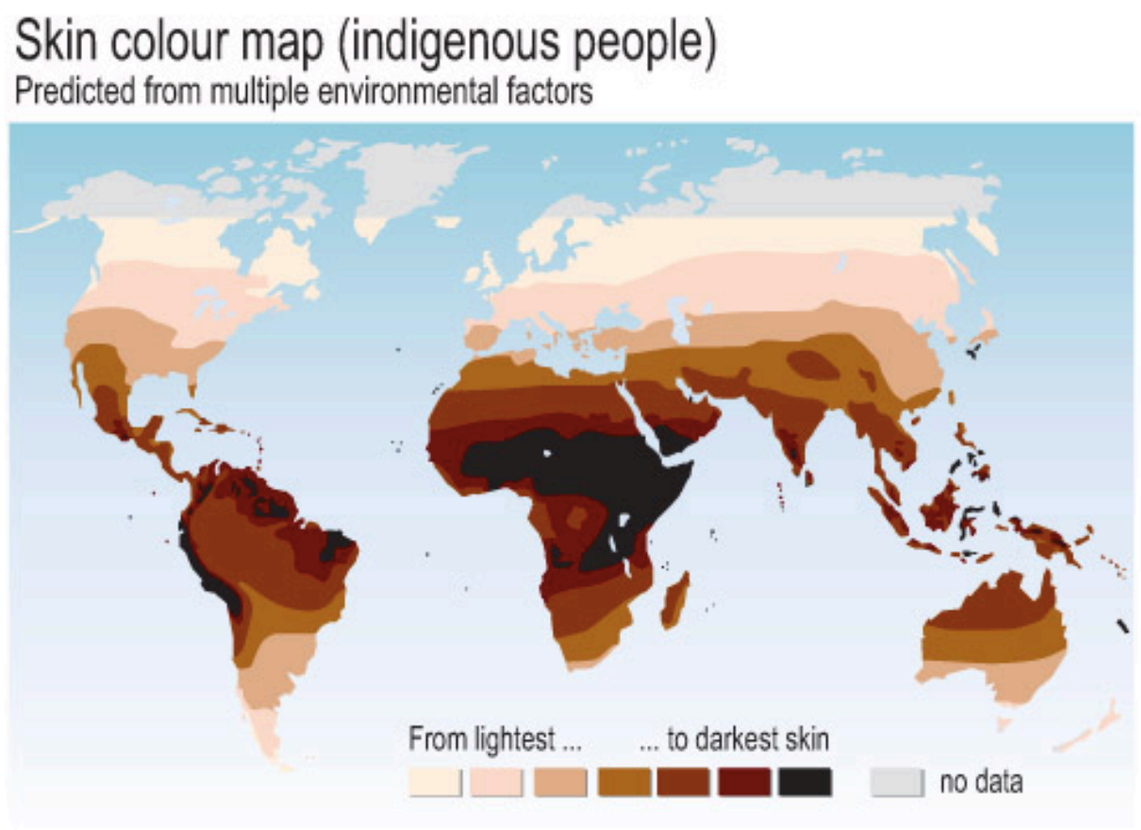
Fig. 9.18 (a) Continuous variation in human height. (b) Bell-shaped curve typical of continuous variation in a trait.

http://www.carolguze.com/text/442-9-complex_multifactorial_inheritance.shtml


Pygmies
Bostonians
Dinkas
Frequency
32
24
16
8
0
130
140
150
160
170
180
190
200
Height (cm)

Reconciling Mendel and Nature

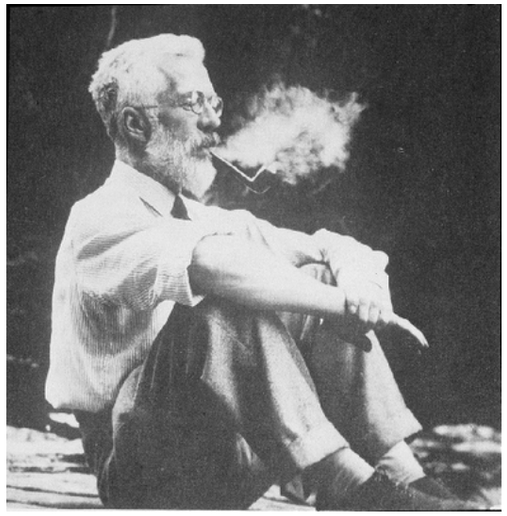
How do Mendelian genes encode continuous variation?



Why aren't there 27 shades of human skin?

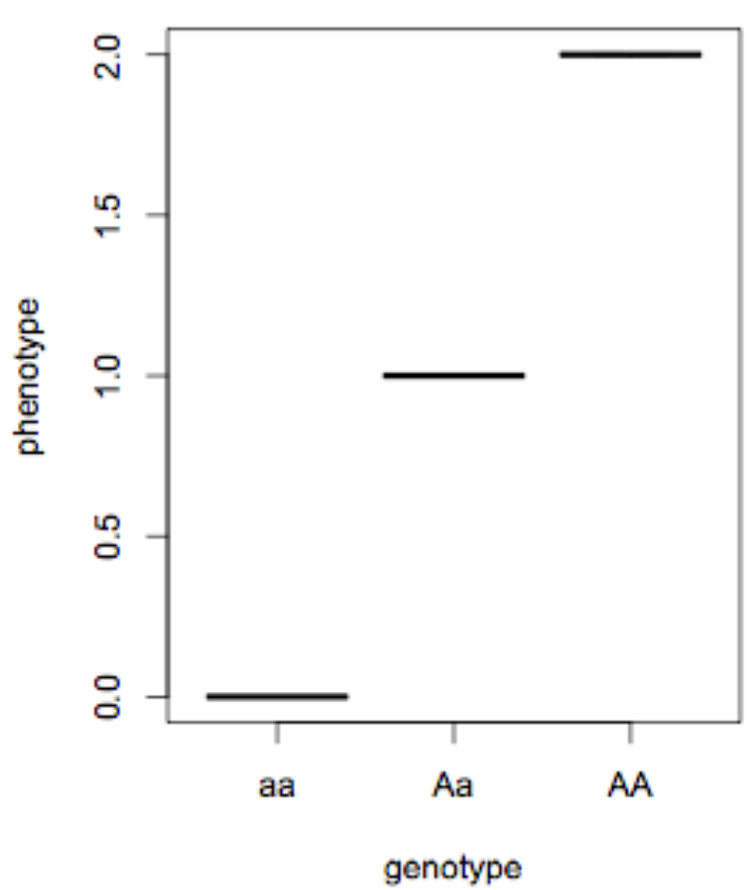
Reconciling Mendel and Nature

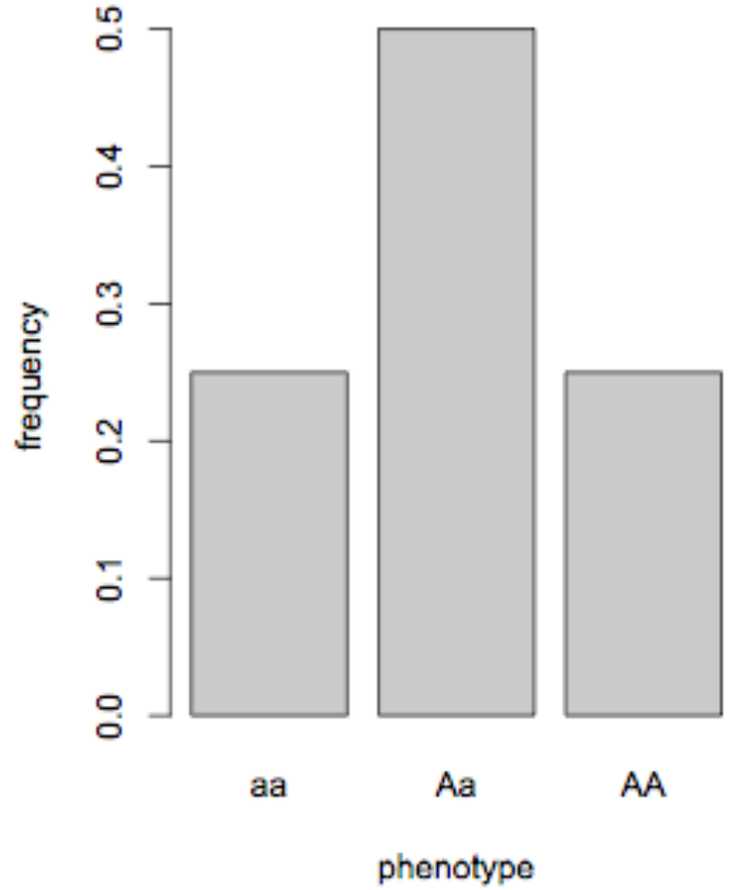
How do Mendelian genes encode continuous variation?



R.A. Fisher

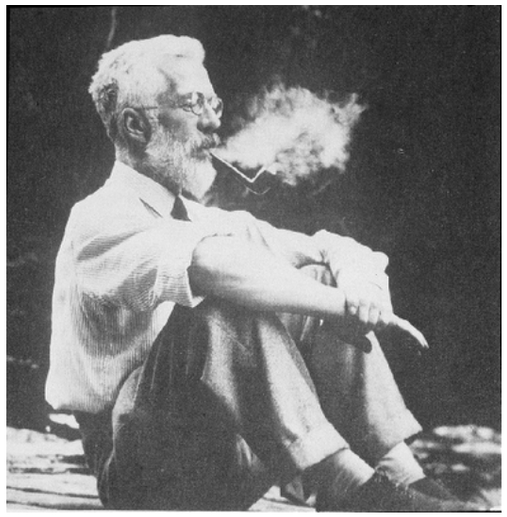
One locus trait, each A allele adds one "unit" of phenotype
Cross: Aa x Aa





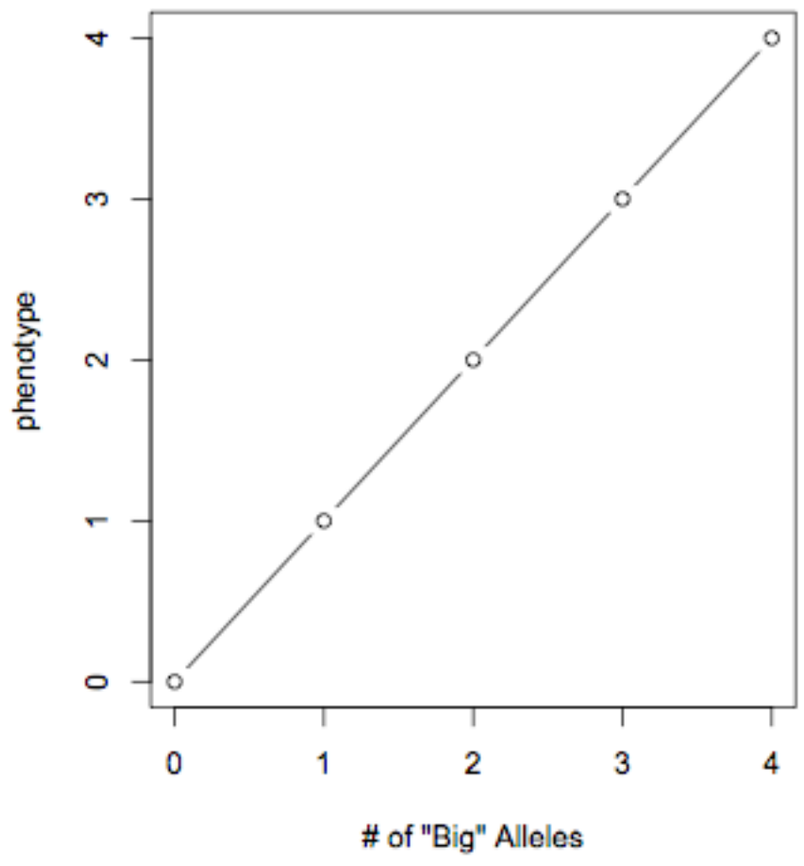
Reconciling Mendel and Nature

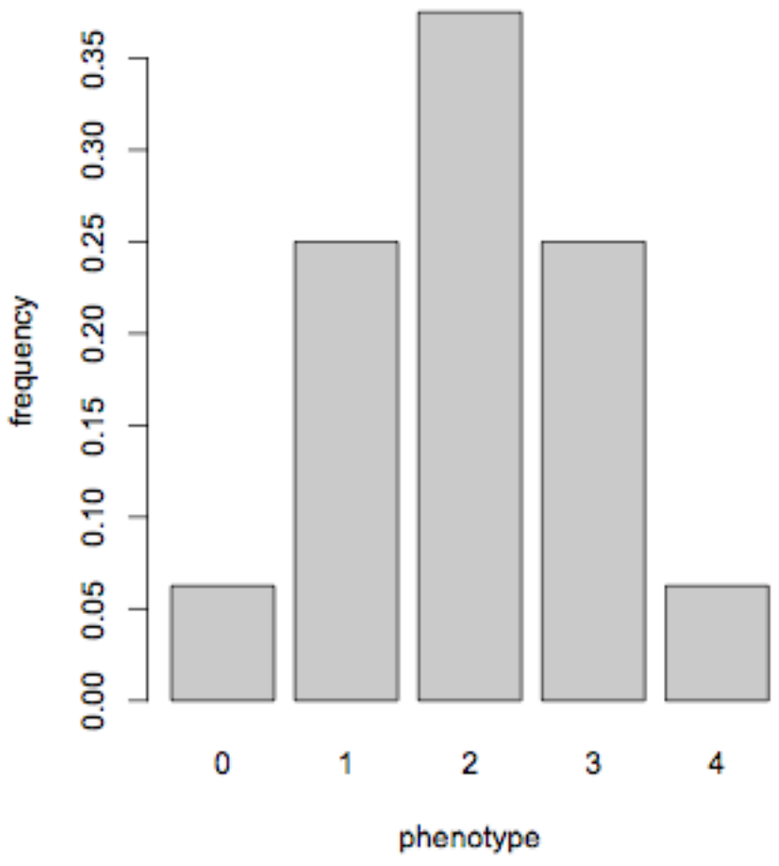
How do Mendelian genes encode continuous variation?



R.A. Fisher

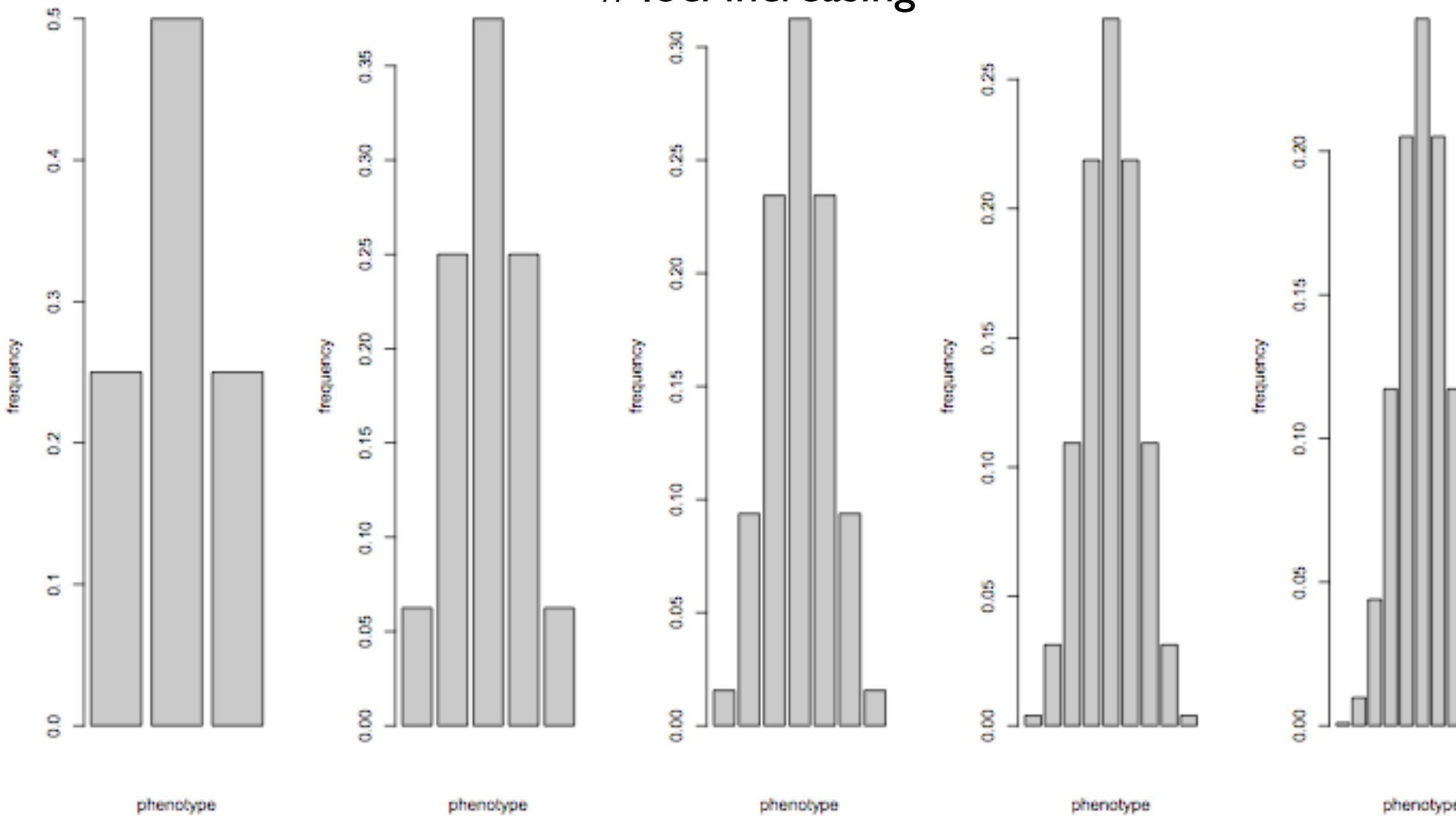
Two locus trait, cross: AaBb x AaBb





Reconciling Mendel and Nature

loci increasing ⟶



Central Limit Theorem

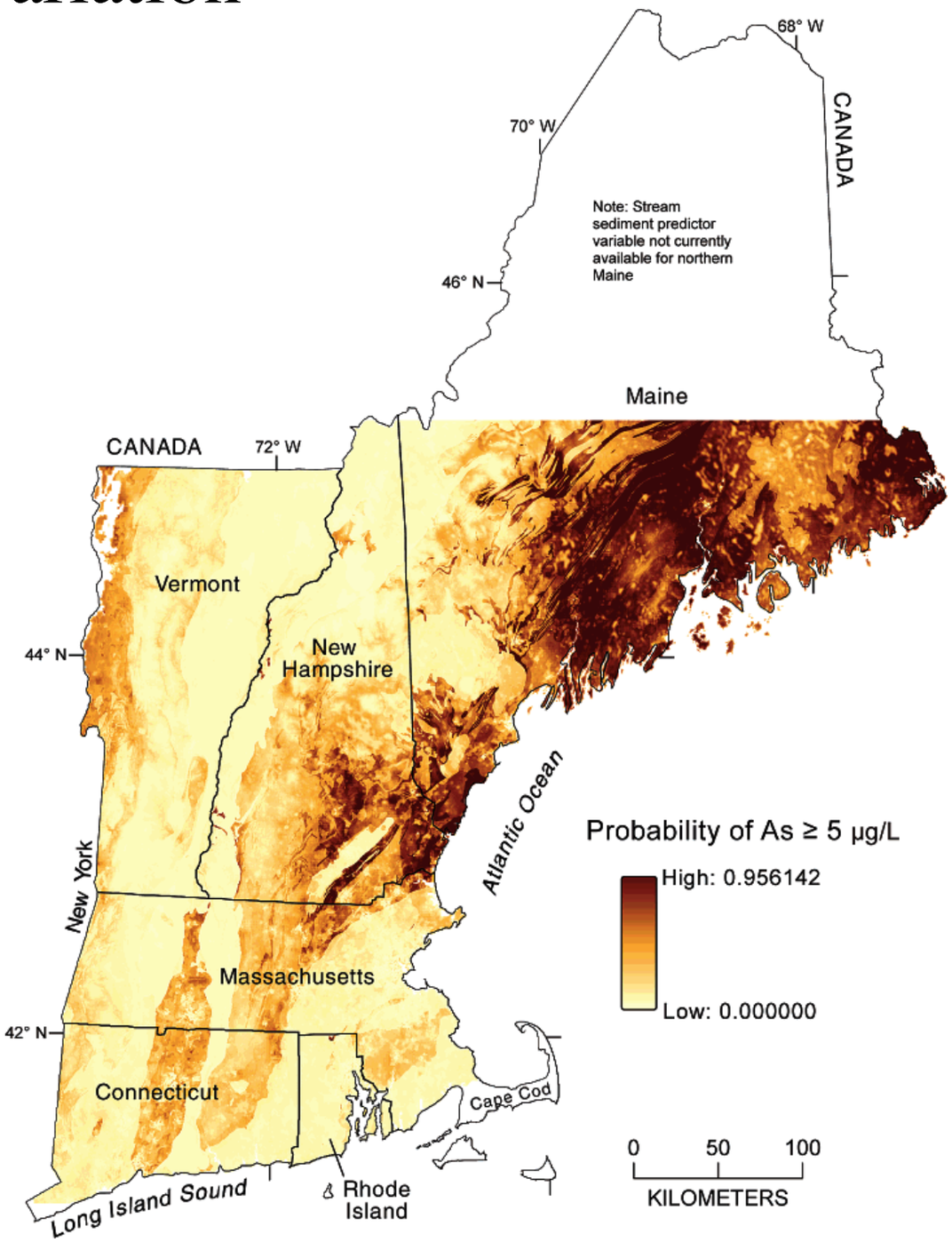
Sources of phenotypic variation?

Genetic Variation

Sources of phenotypic variation?

Environmental Variation

light shade

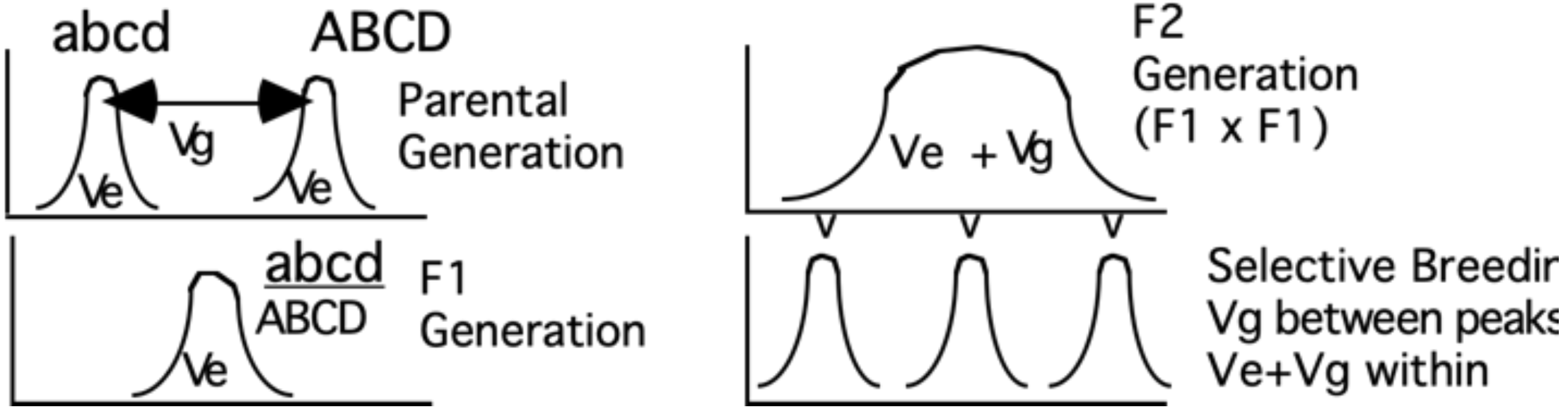


Sources of phenotypic variation?

Variance Components

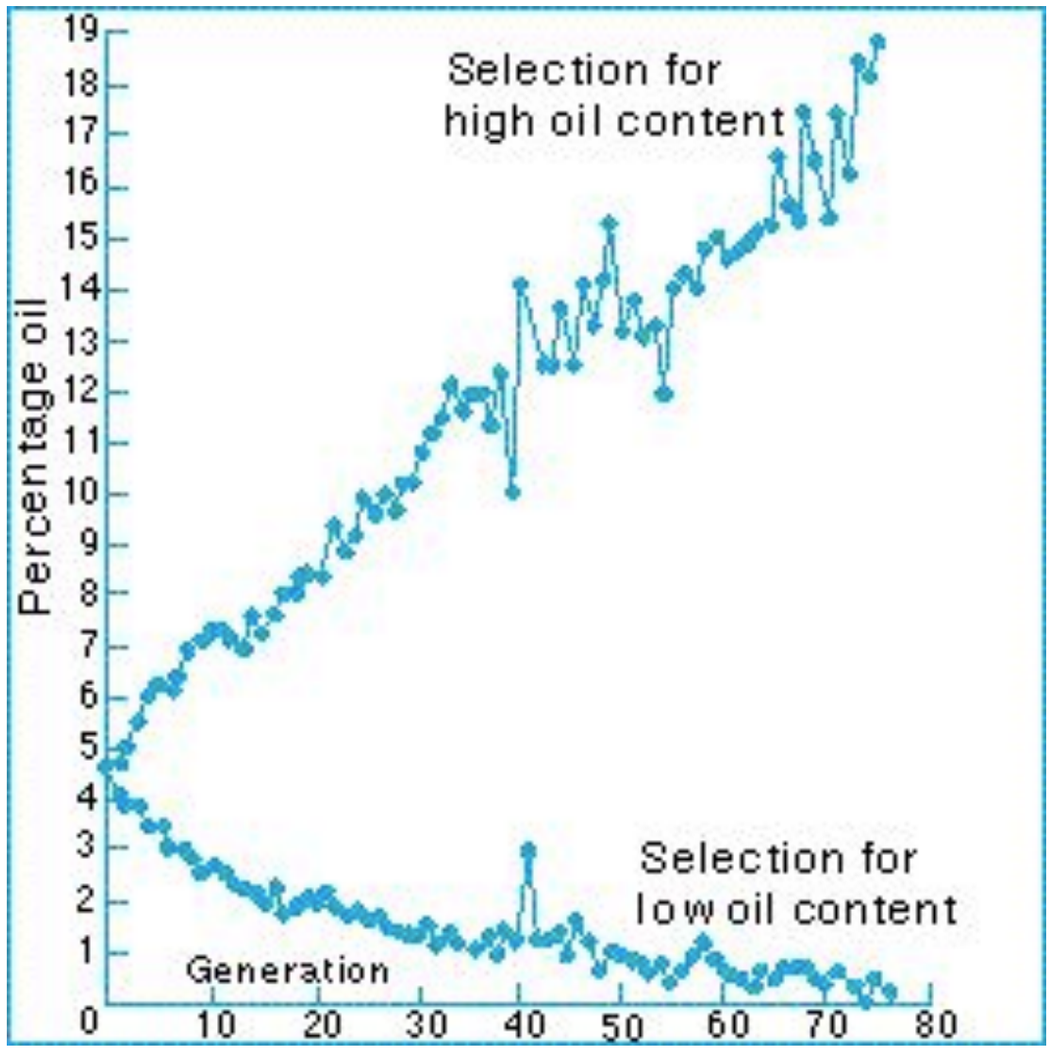
Phenotypic Variation = Genetic Variation + Environmental Variation

$V_p = V_g + V_e$

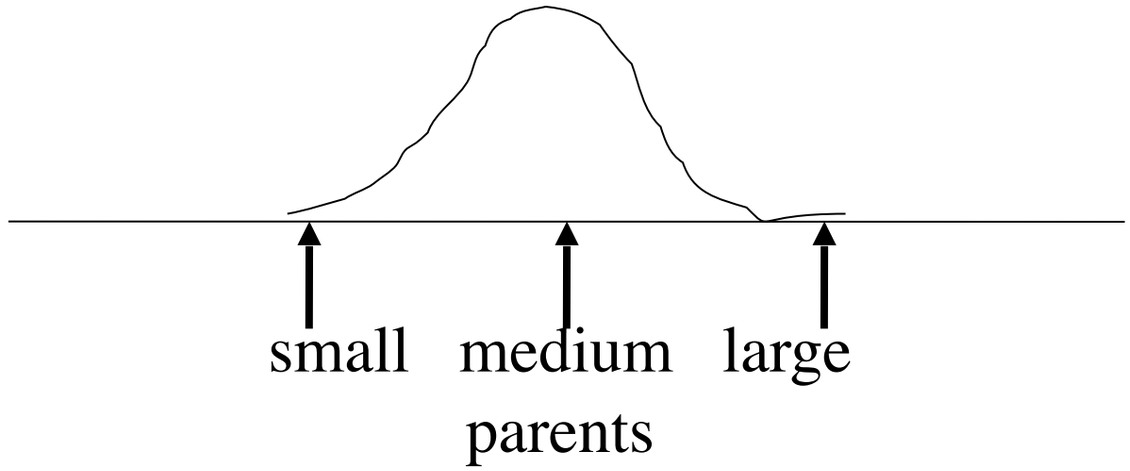


Sources of phenotypic variation?

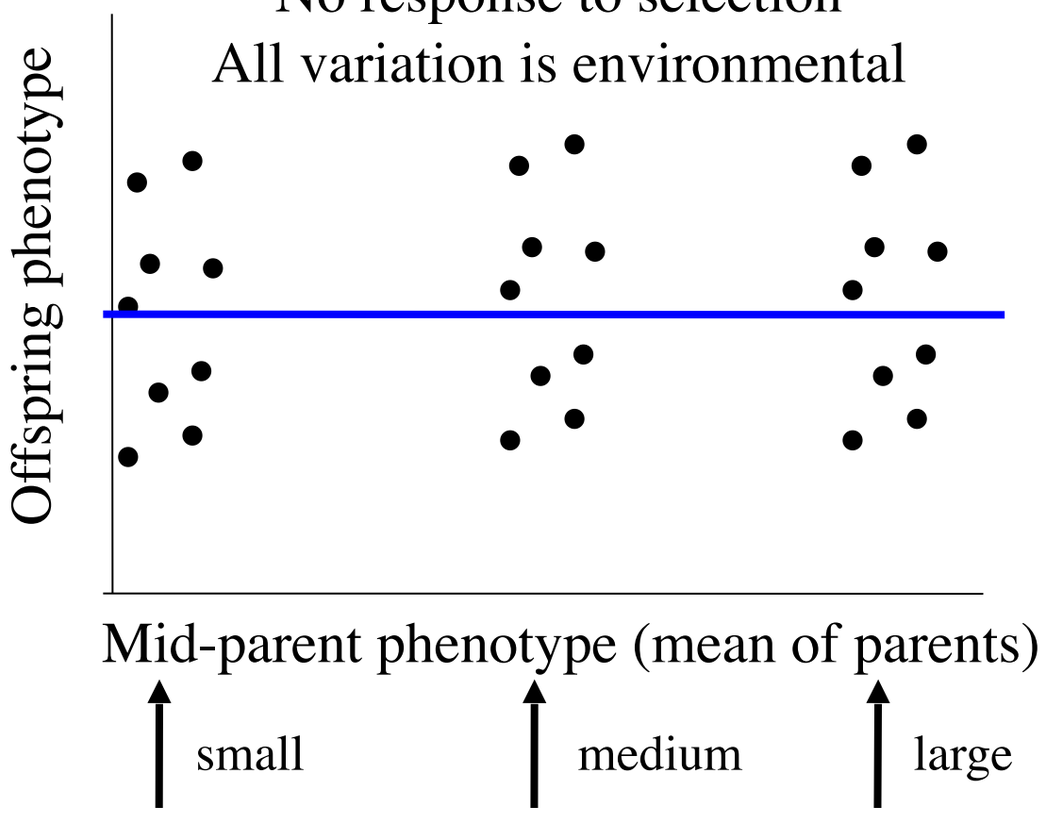
Heritability



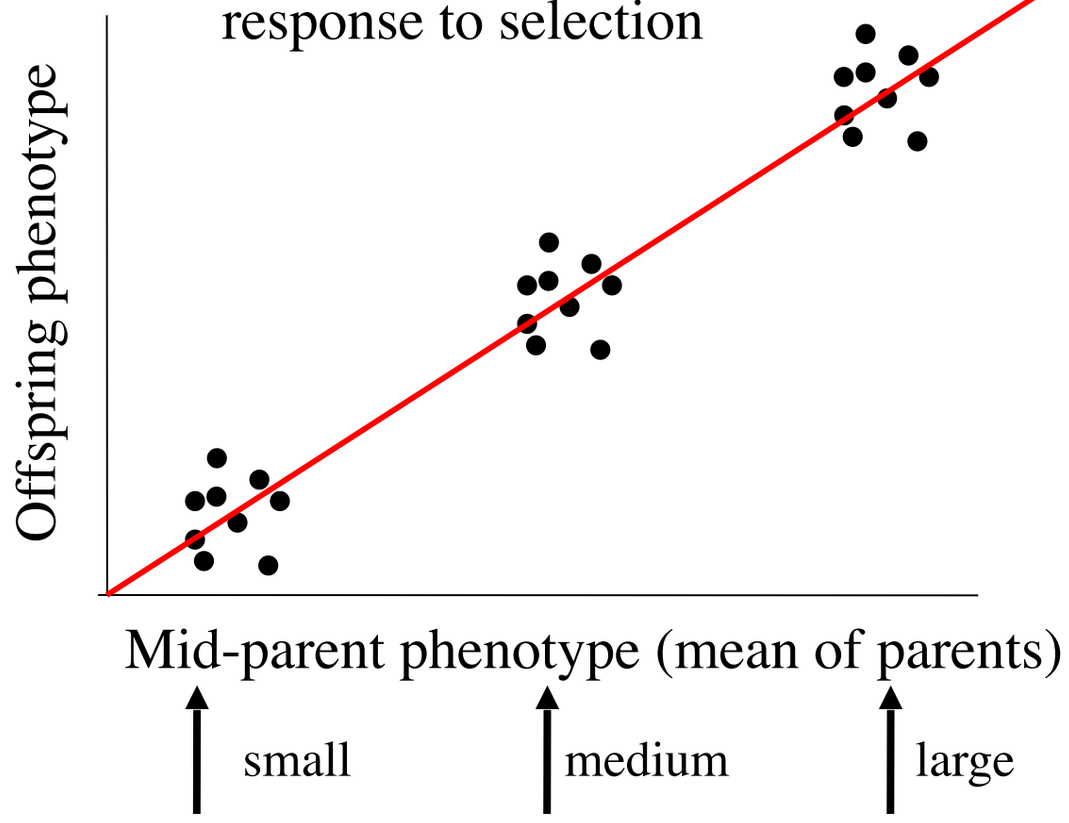
Genetic variation, environmental variation and heritability



Without genetic variation:
No response to selection
All variation is environmental

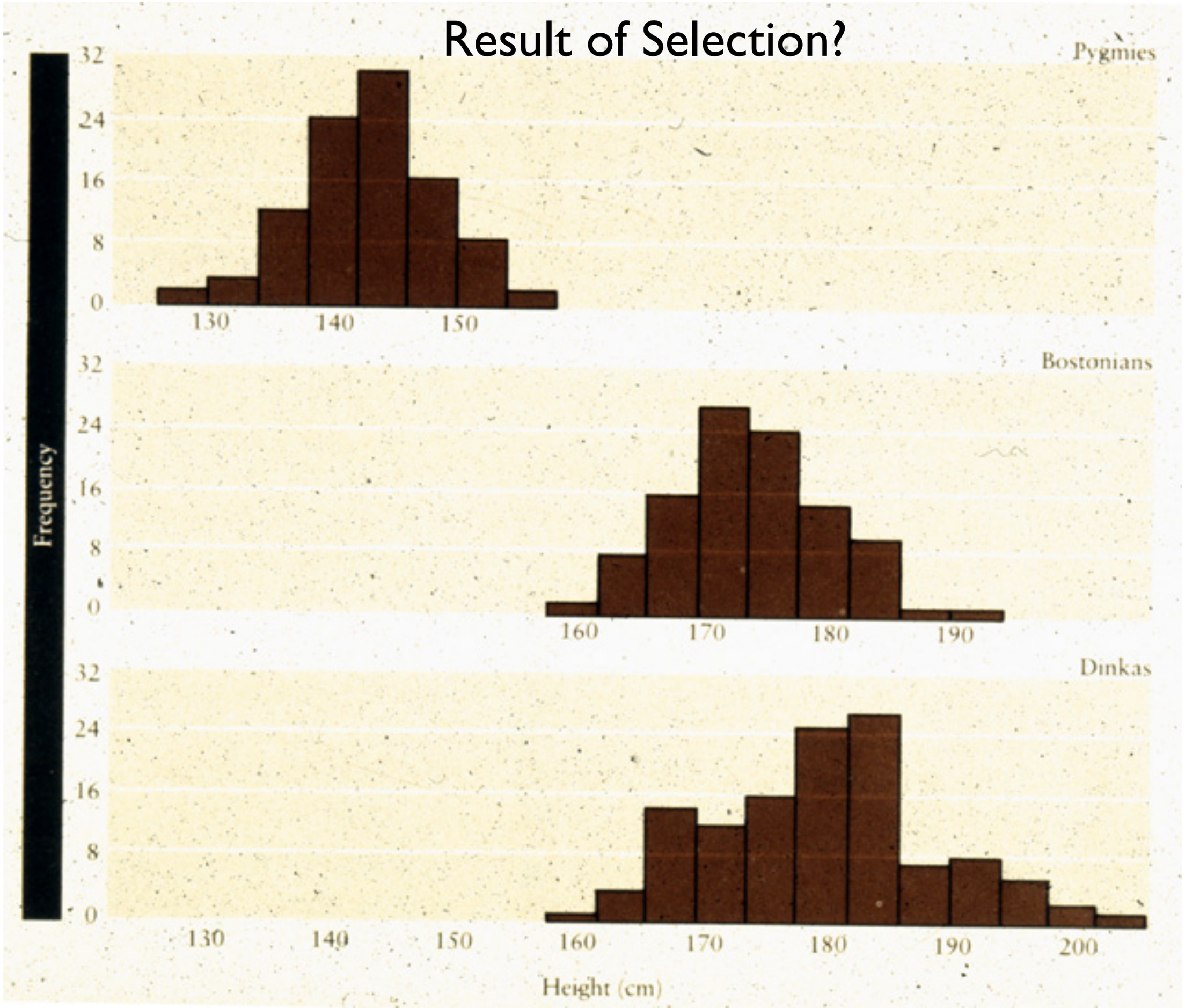


With genetic variation:
response to selection



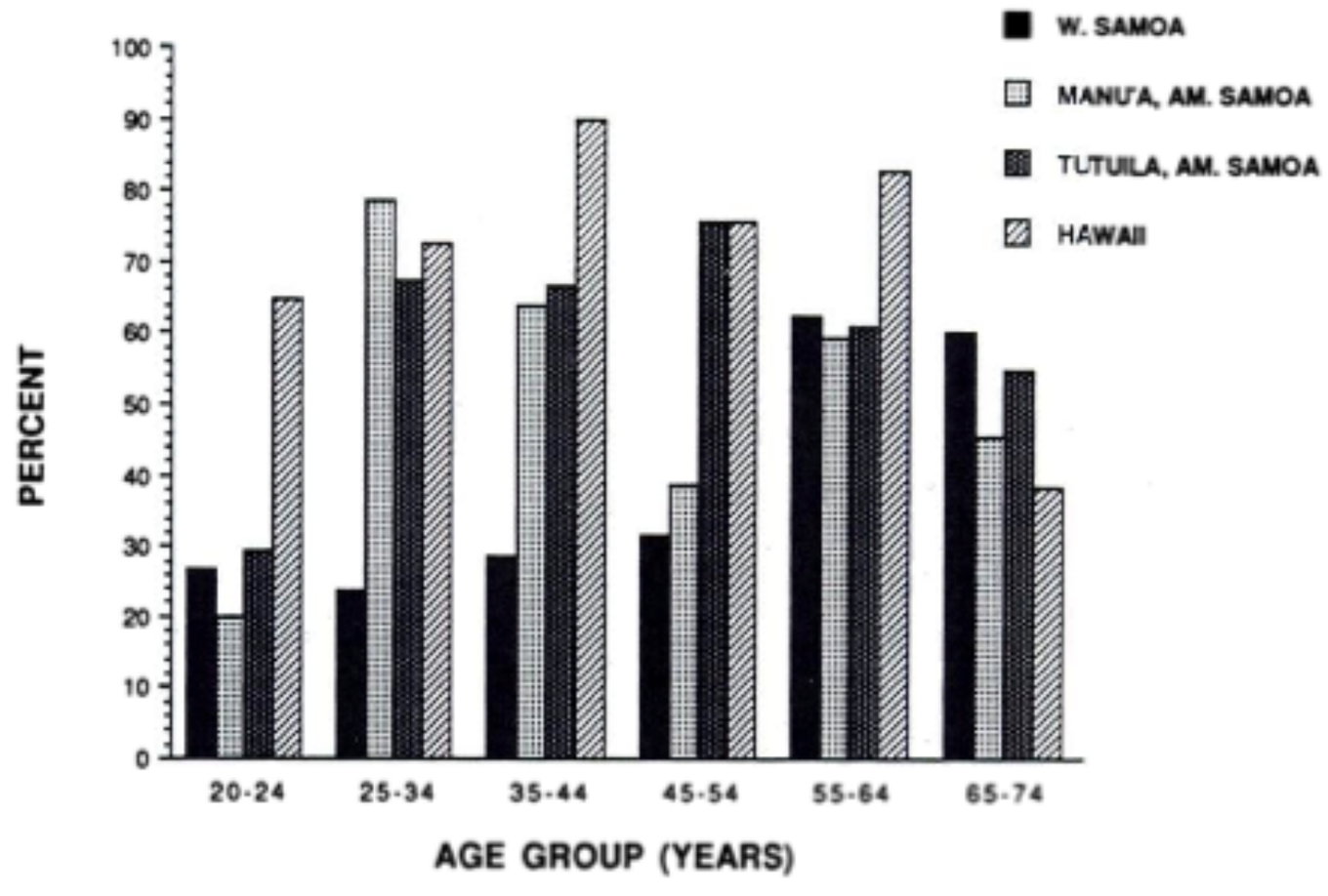
Heritability = V_G / V_P = Slope of Parent Offspring regression

Result of Selection?



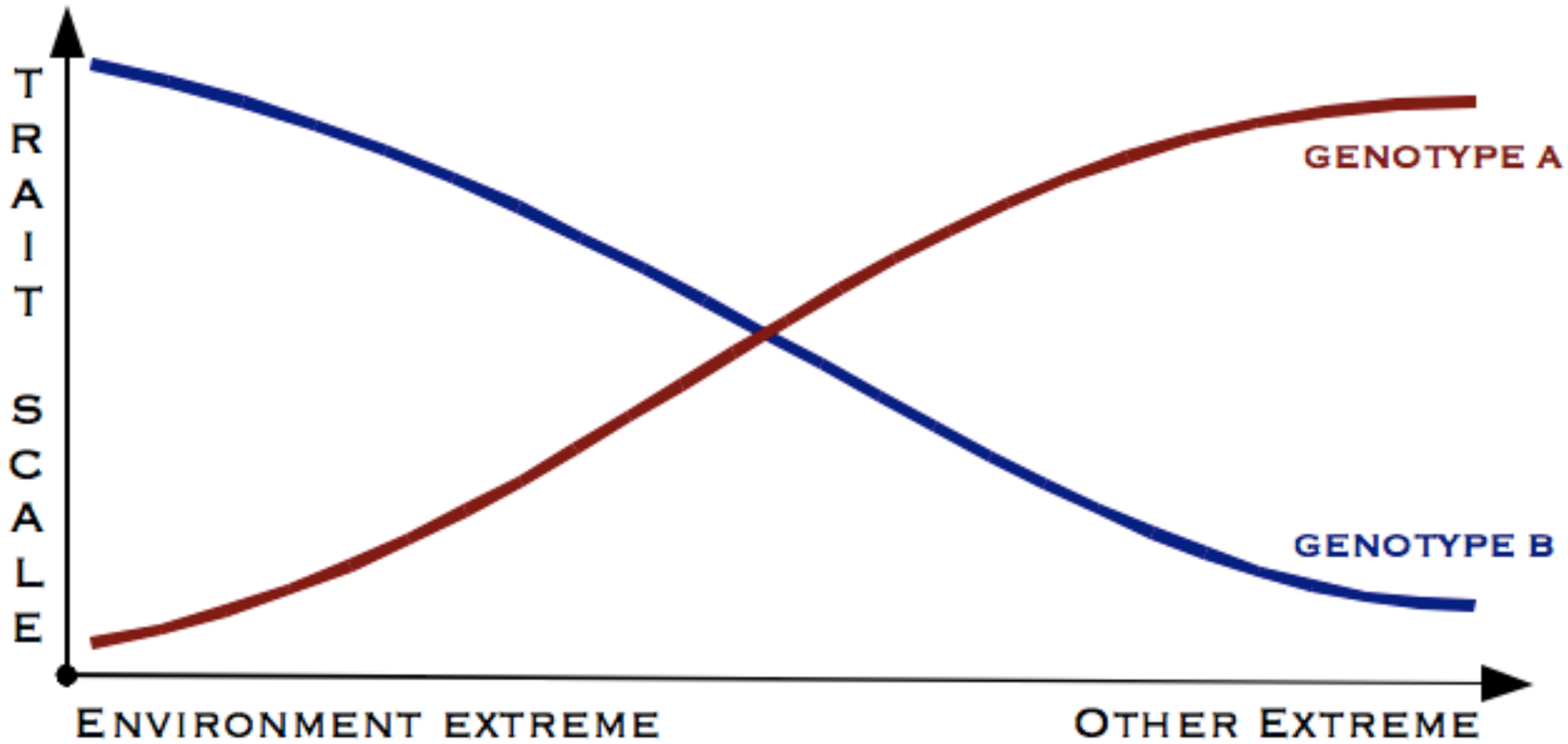
Sources of phenotypic variation?

Gene by Environment Interaction
(a.k.a. GxE)



Sources of phenotypic variation?

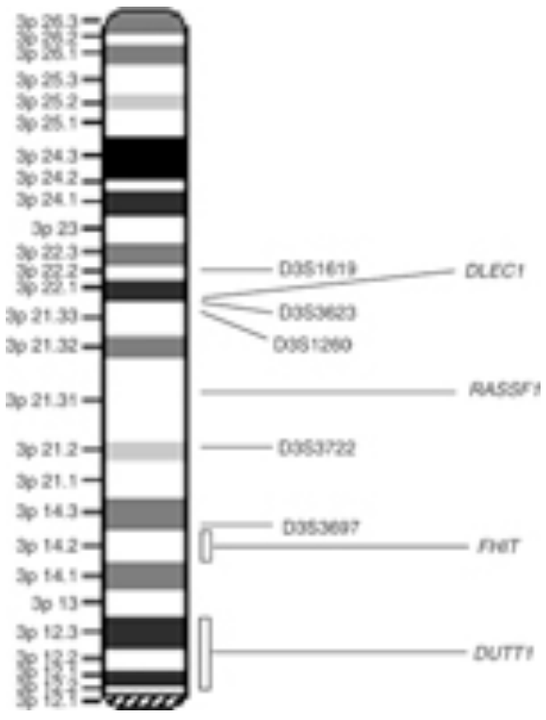
Gene by Environment Interaction
(a.k.a. GxE)

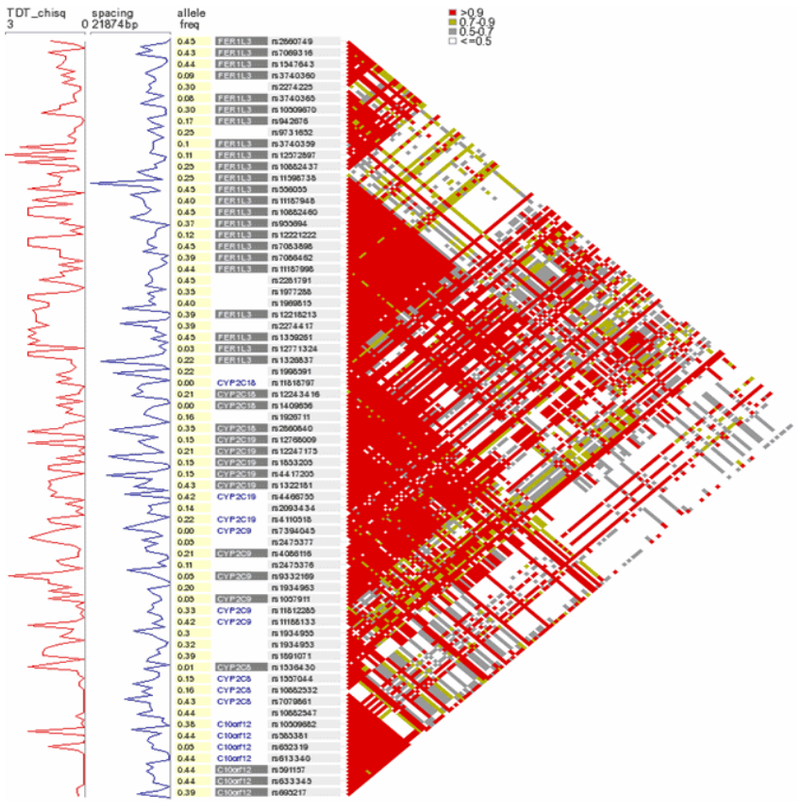


Differential response among genotypes to environmental gradient

Mapping Genes

Markers ⟶ Linkage ⟶ Quantitative Phenotype





A living histogram. These students and faculty at the University of Connecticut have sorted themselves into columns by height. (Peter Morenus, University of Connecticut)

Want to find genetic variant(s) which explain genetic variation underlying a trait / disease

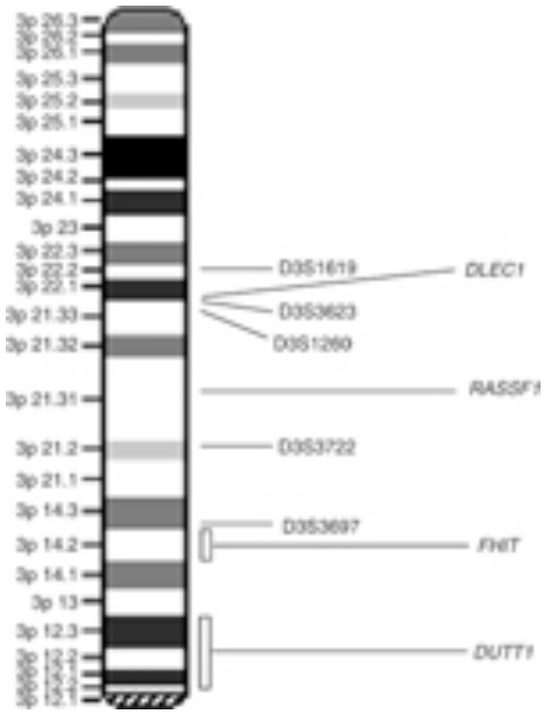
Freeman and Herron 2001.

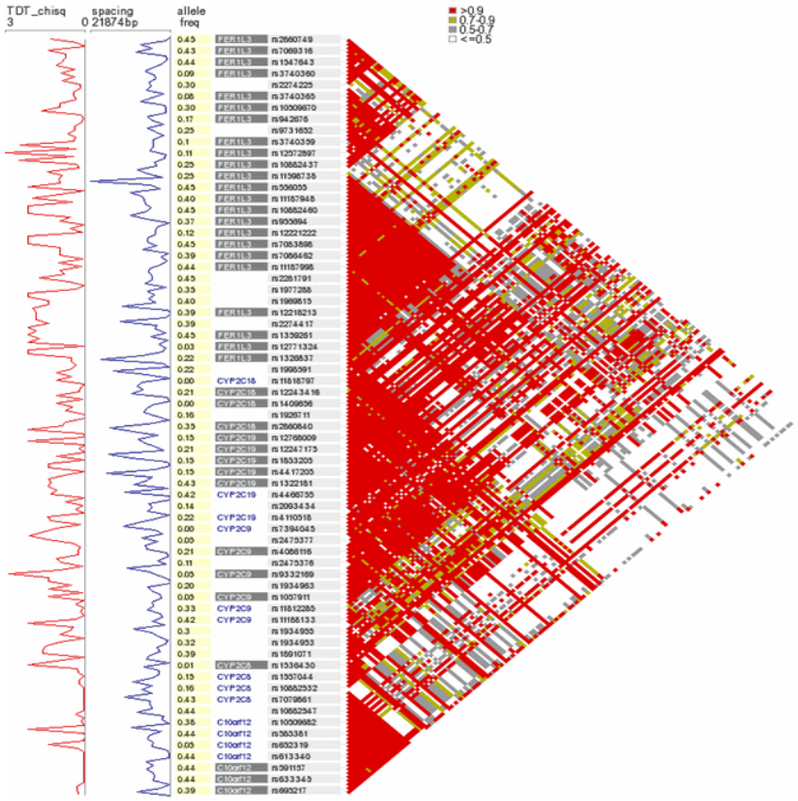
http://www.mrc.gm/images/Genetics_Analysis_of_Pattern.gif

http://www.nature.com/bjc/journal/v90/n10/full/6601775a.html

Mapping Genes

Markers ⟶ Linkage ⟶ Case-Control





Want to find genetic variant(s) which explain genetic variation underlying a trait / disease

Freeman and Herron 2001.

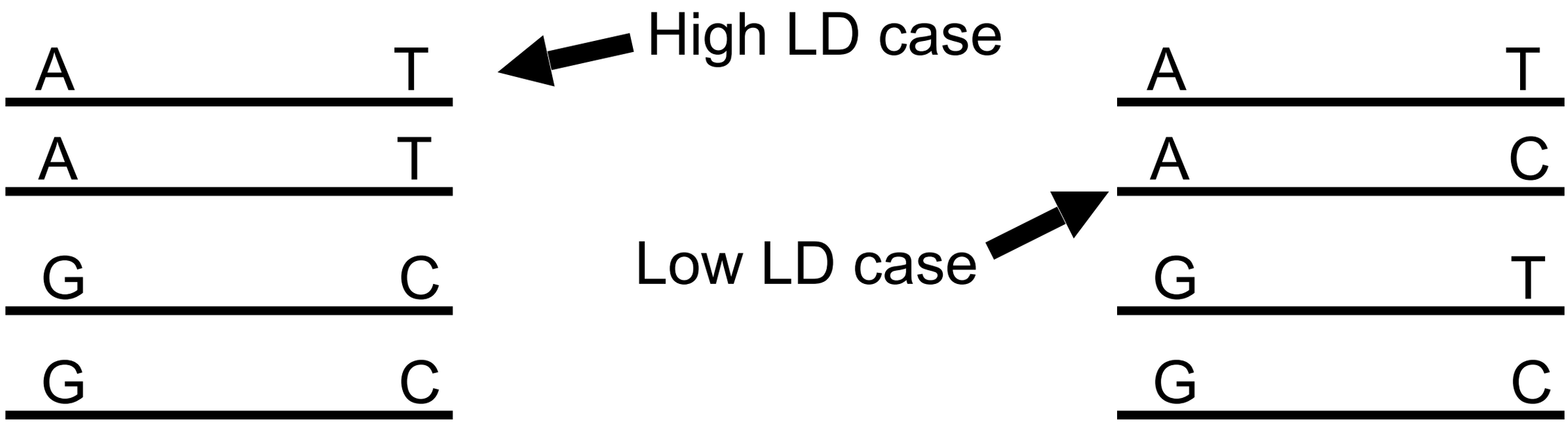
http://www.mrc.gm/images/Genetics_Analysis_of_Pattern.gif

http://www.nature.com/bjc/journal/v90/n10/full/6601775a.html

Quantifying Linkage

Linkage: the tendency for traits to be co-inherited.
results from physical organisation of chroms.

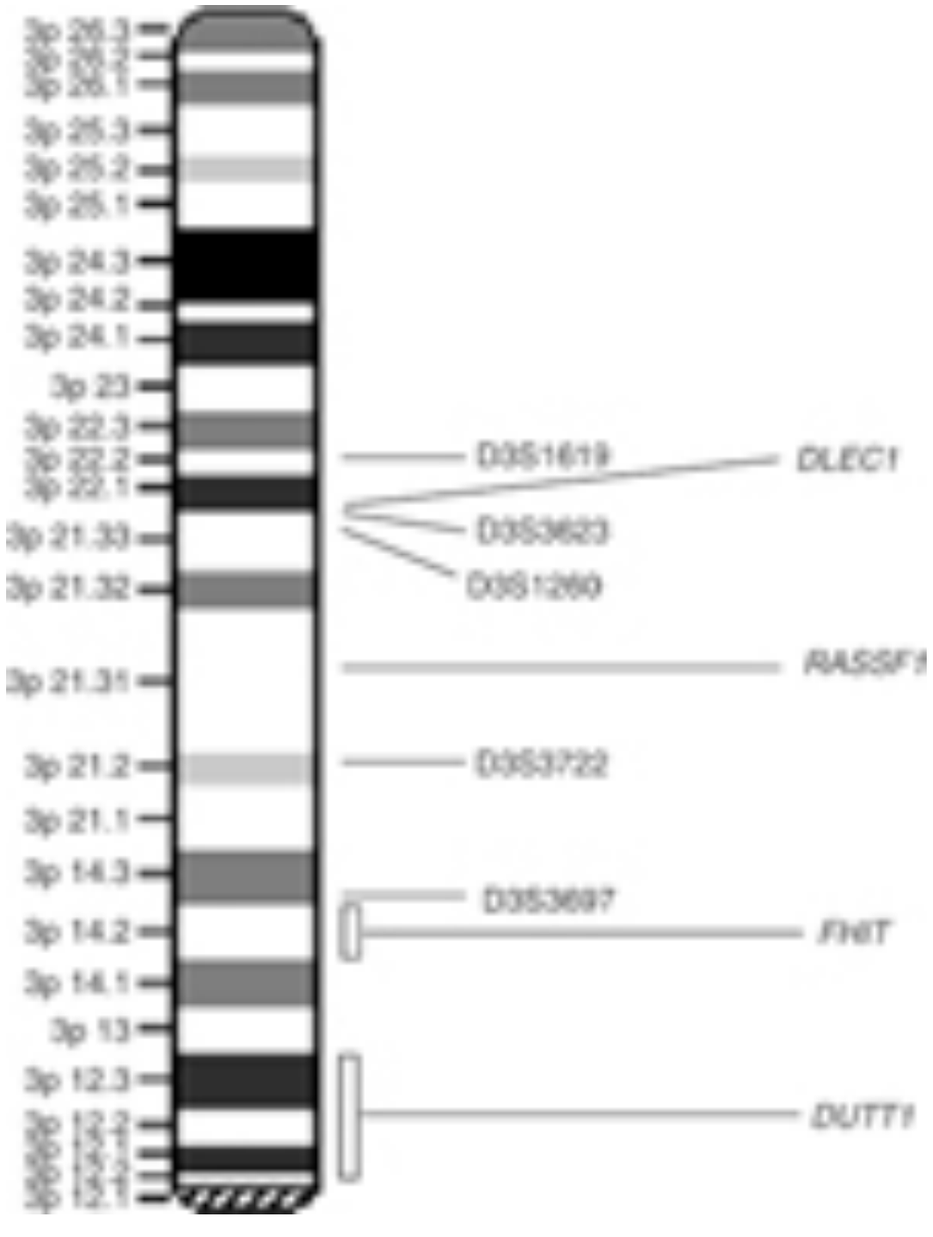
Linkage Disequilibrium: (LD) “amount” by which alleles are correlated (or covary) in a population.

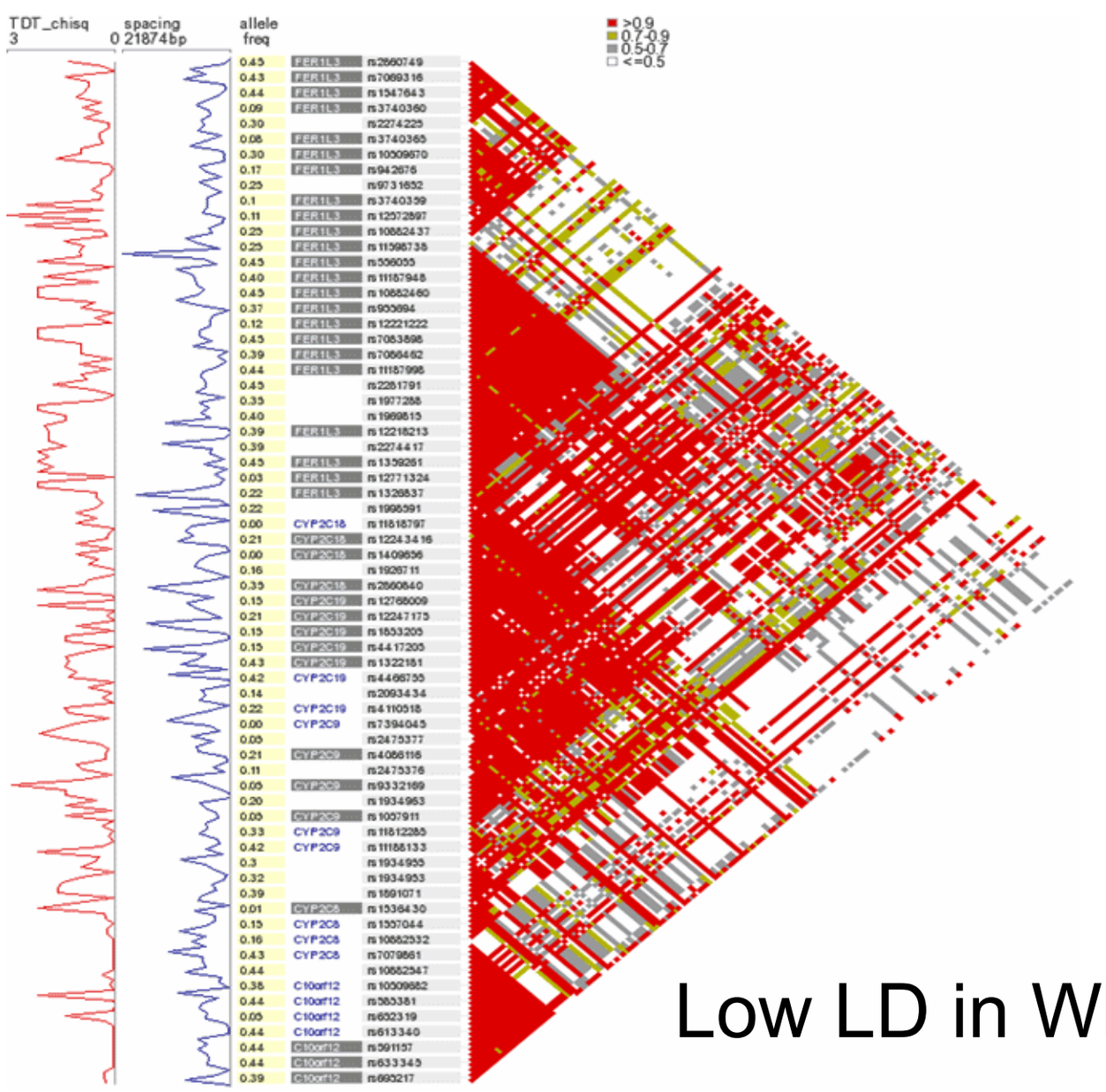


Quantifying Linkage

Pairwise LD comparison matrix

High LD in Red

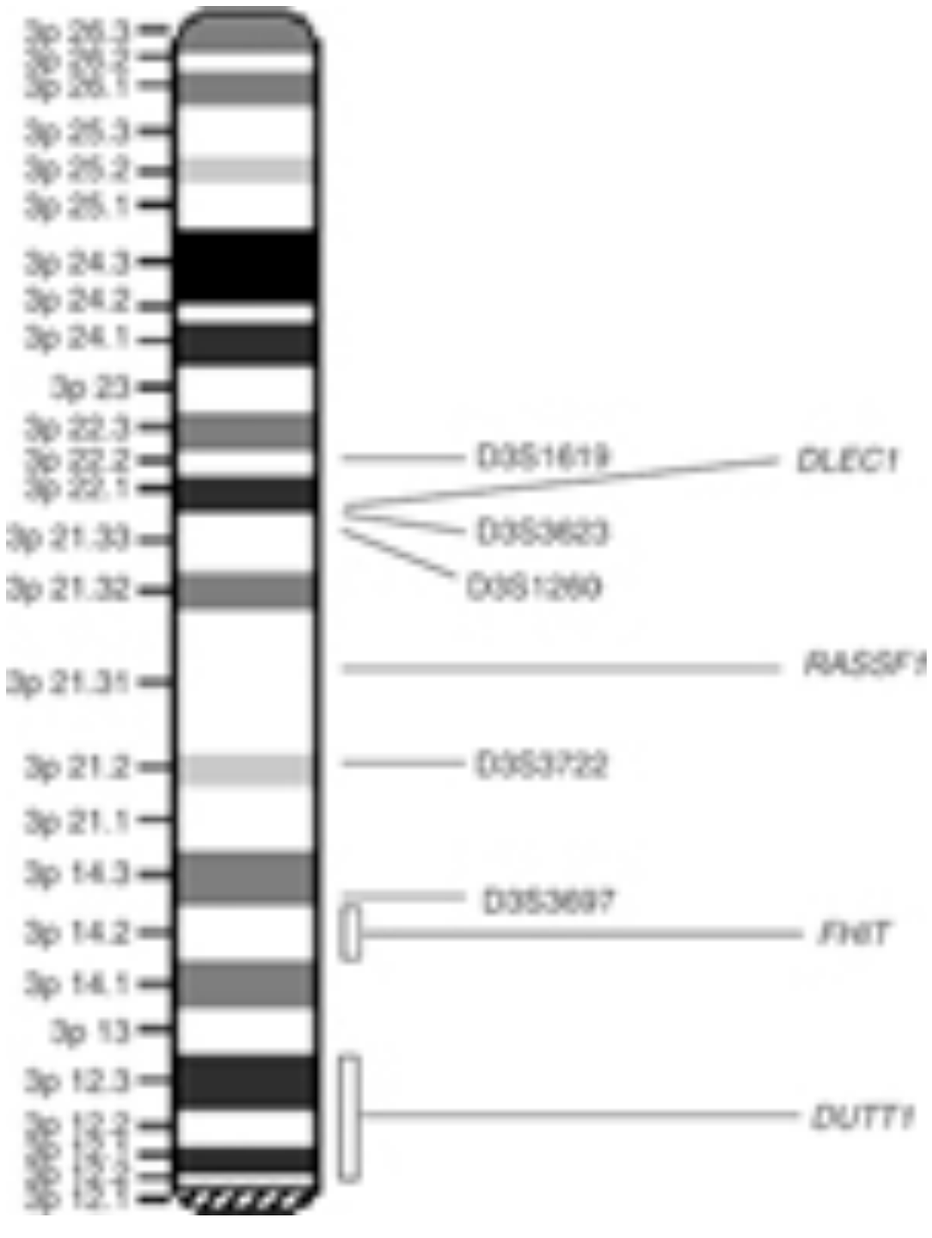


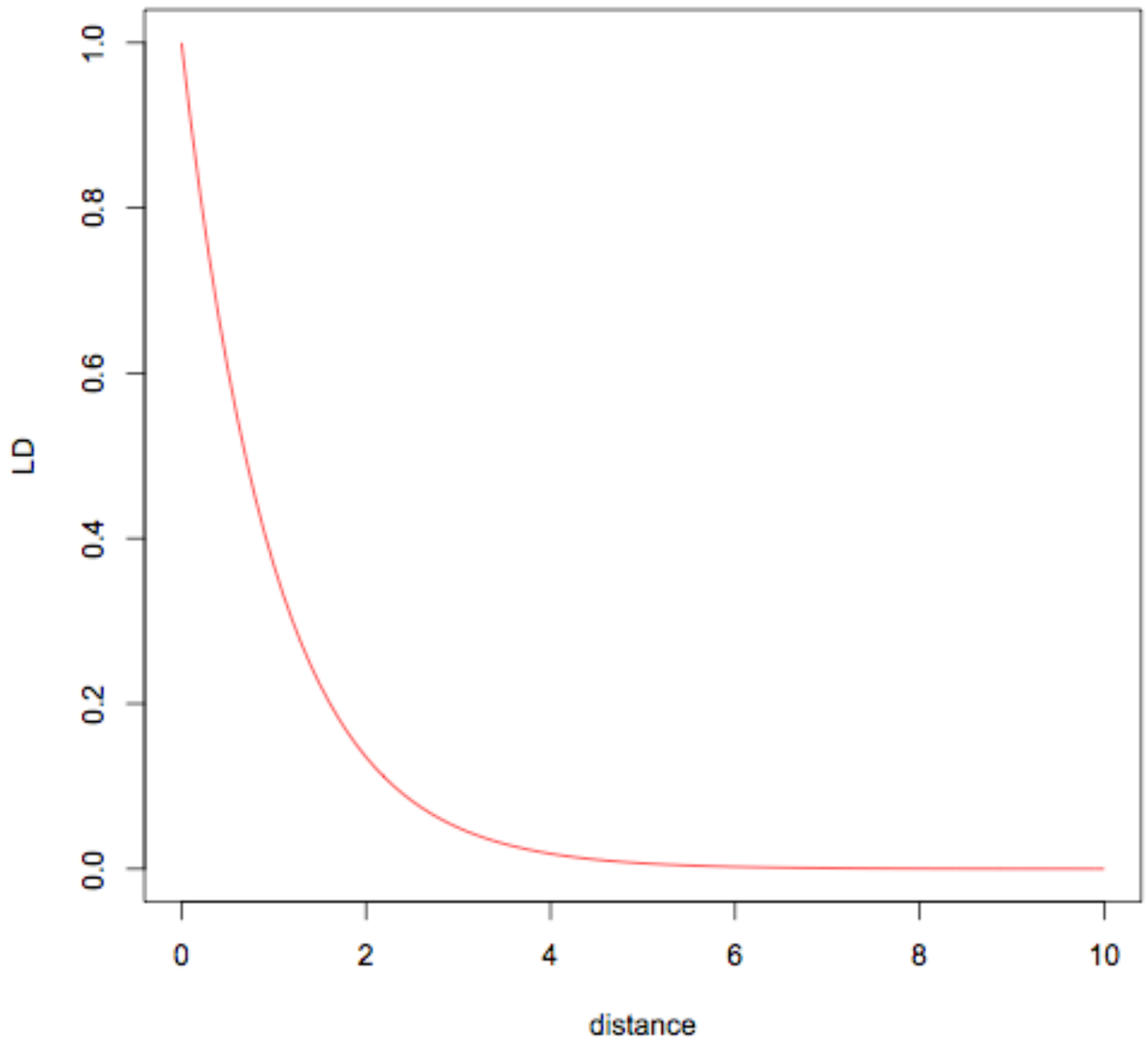


Low LD in White

Quantifying Linkage

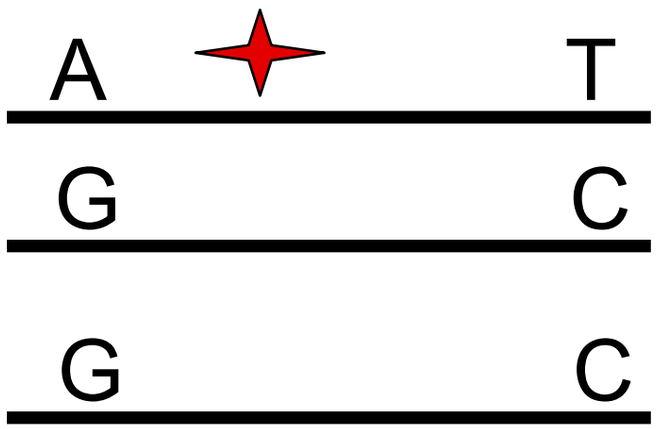
LD decays with distance





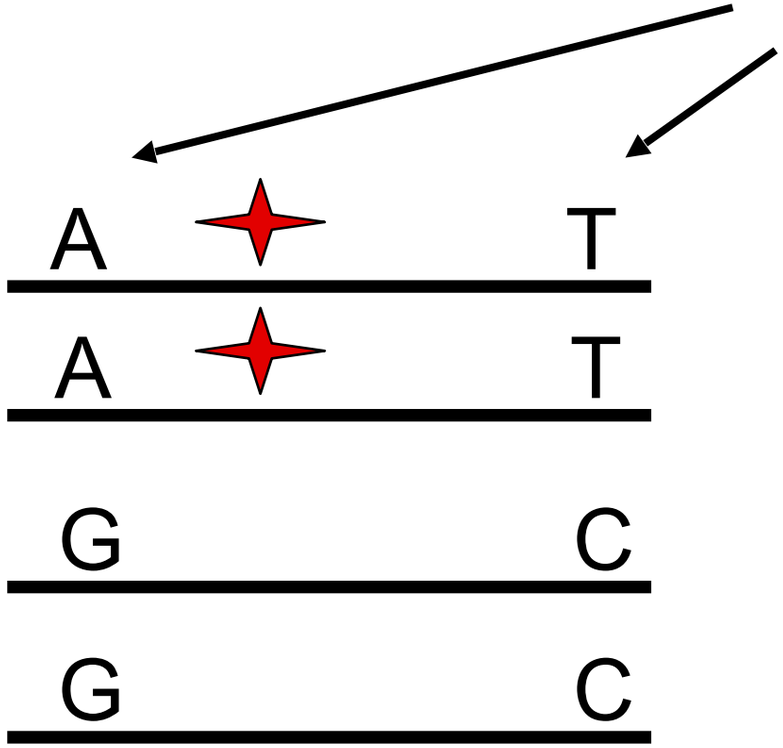
Quantifying Linkage

Intuition behind mapping disease loci

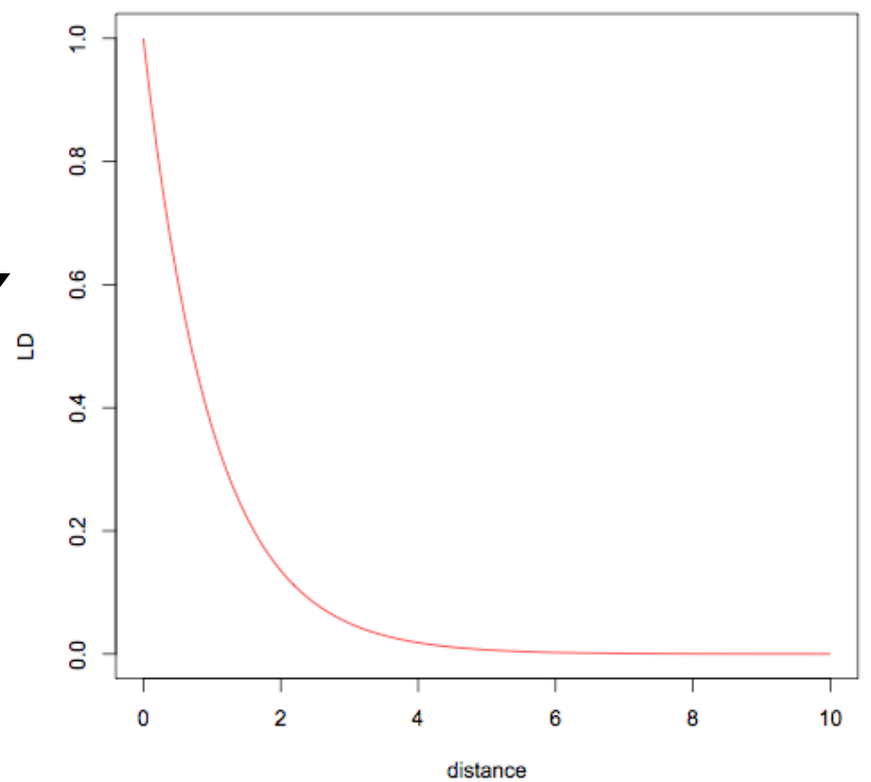


1) Unique origin of disease allele

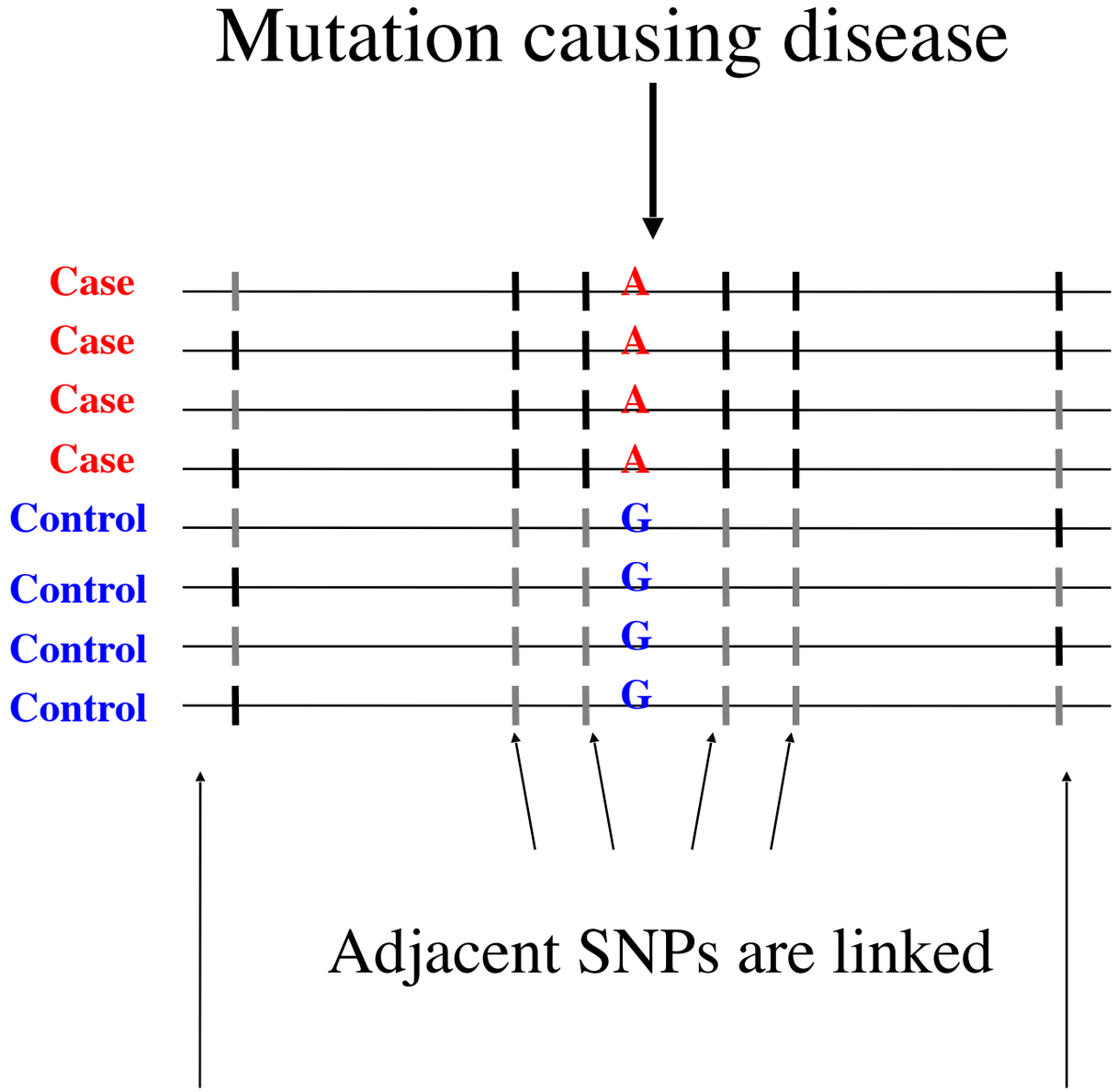
2) High LD locally creates association At non-causitive, marker loci



3) LD decays as we move farther away



Linkage Mapping Overview



Hard Problem- lots of different techniques to try to find causitive SNPs

Most traits in organisms
Show continuous variation

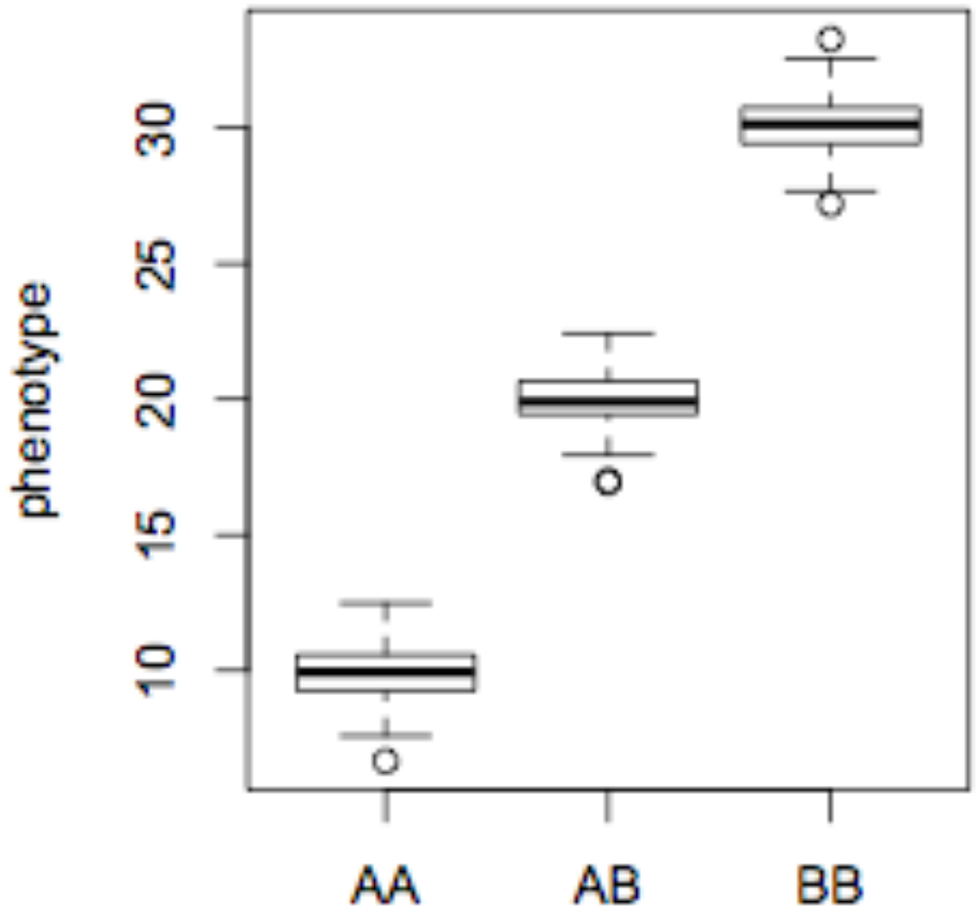
How do we find the genes
That affect these
“quantitative” traits

Scan the genome for
Nucleotide sites that
Co-vary with the
phenotype

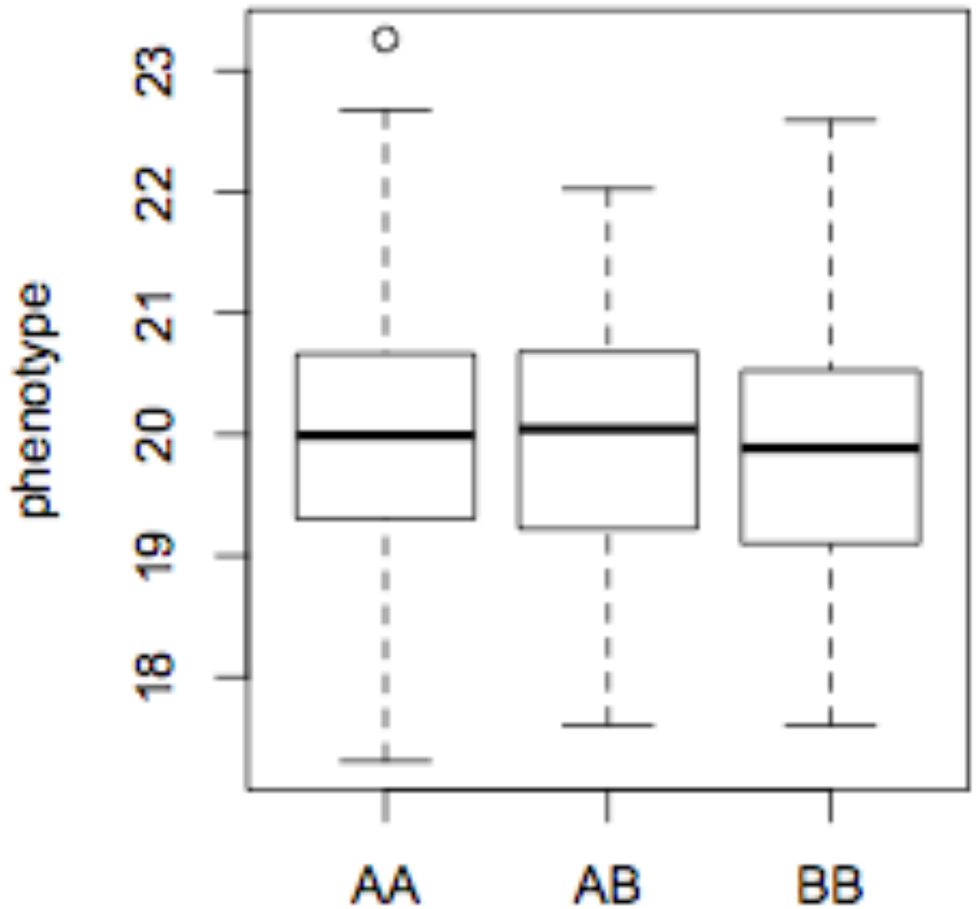
Association Mapping

Go marker my marker and ask:
Is there an association between genotype and phenotype?

Associated Marker

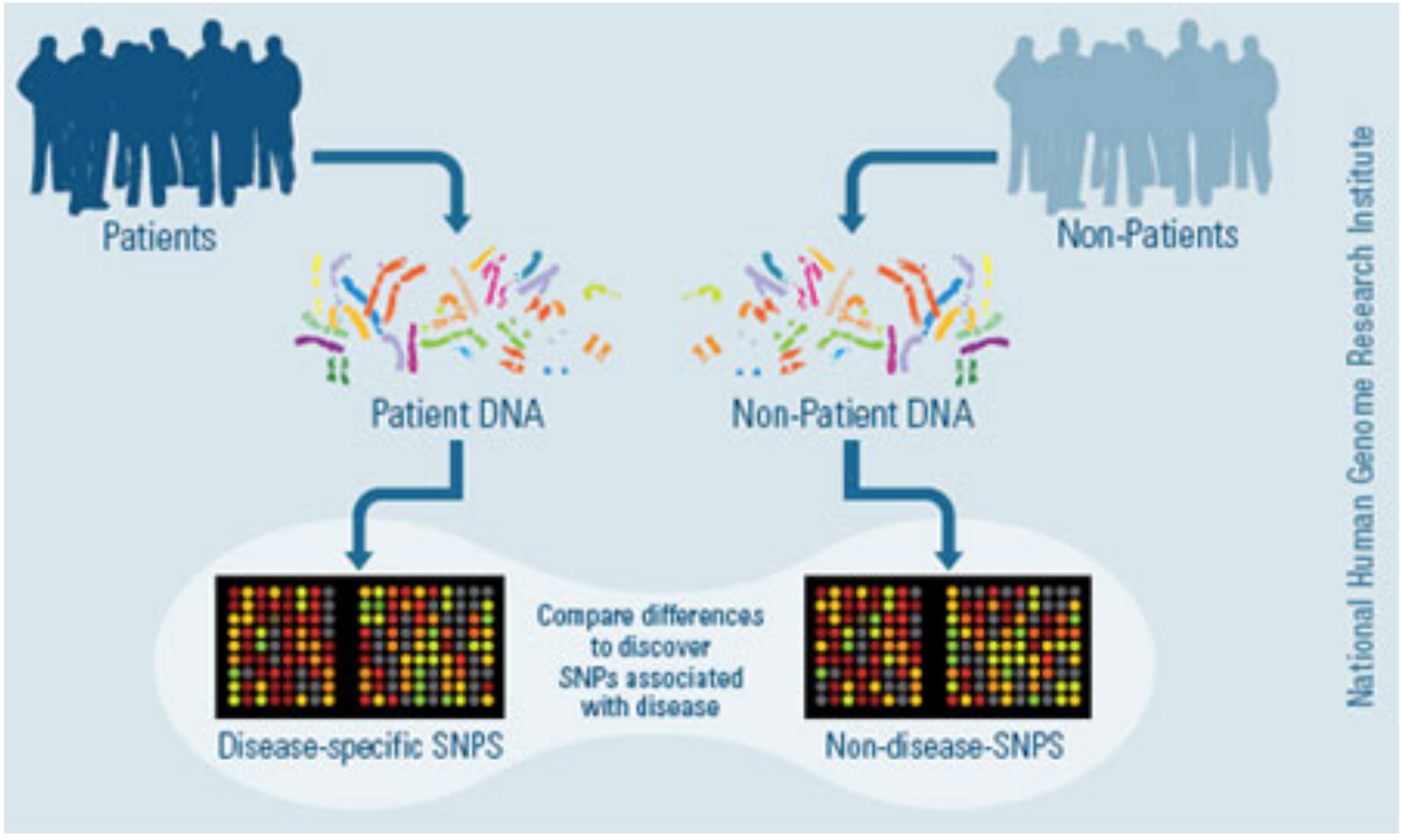


Non Associated Marker



Genome-Wide Association Study

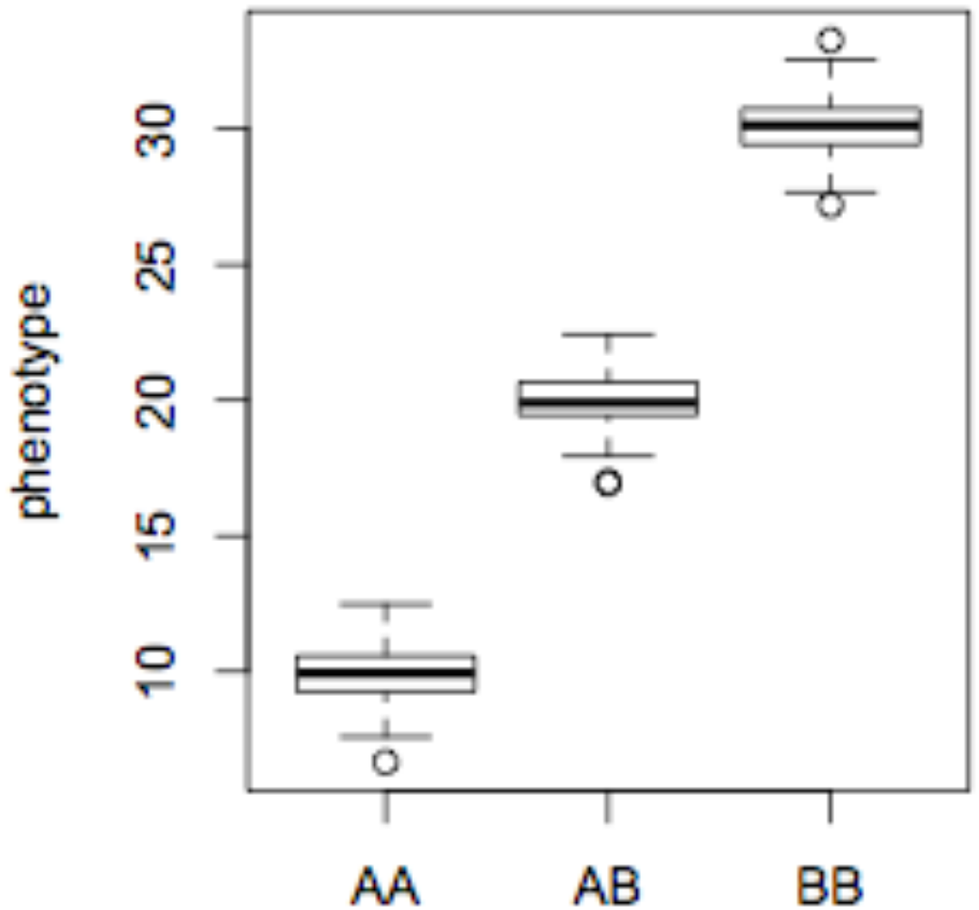
Microarray's allow one to SNP genotype 1.8 million loci in Single experiment



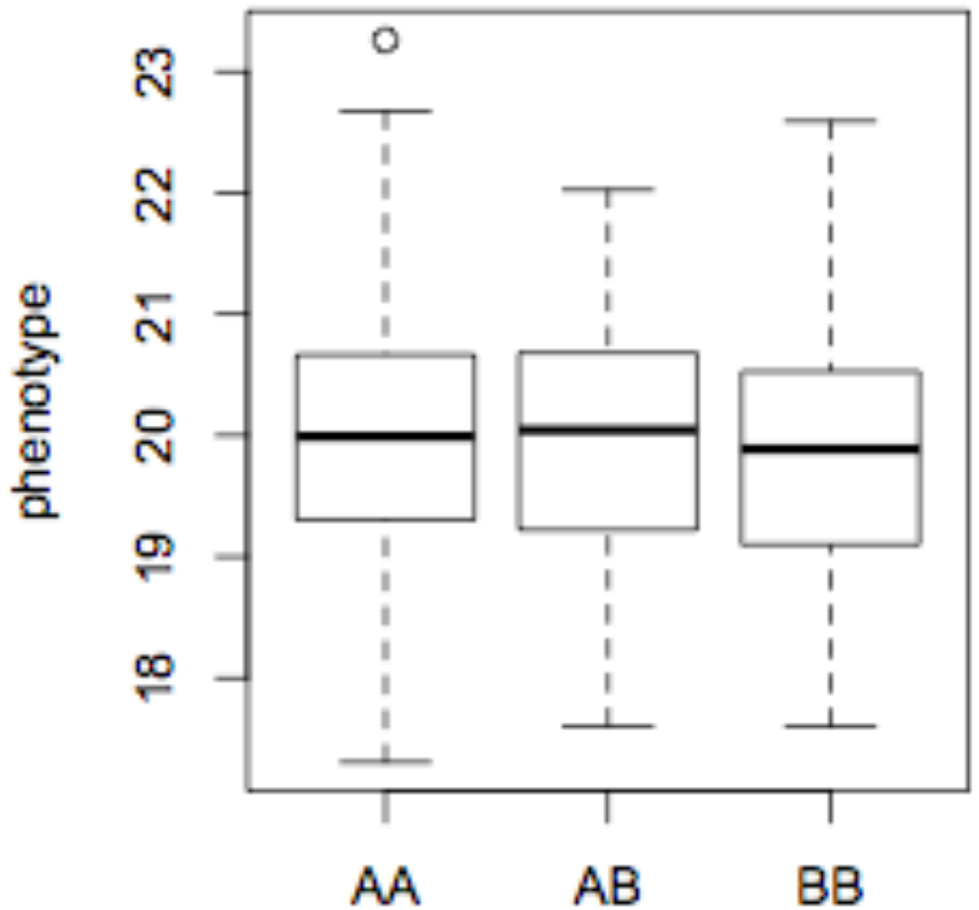
Association Mapping

Use linear regression for this!
pheno = ax + b

Associated Marker



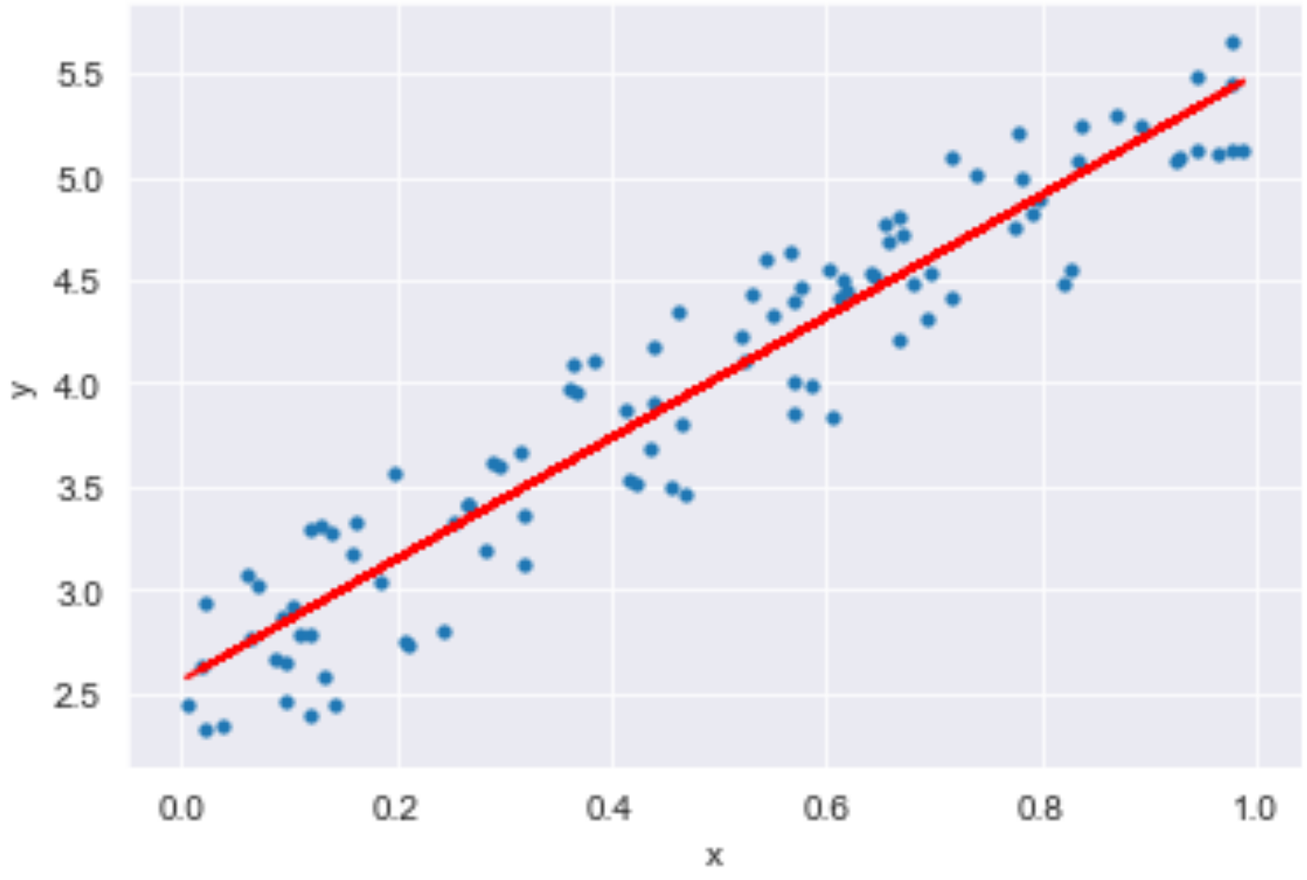
Non Associated Marker



Association Mapping

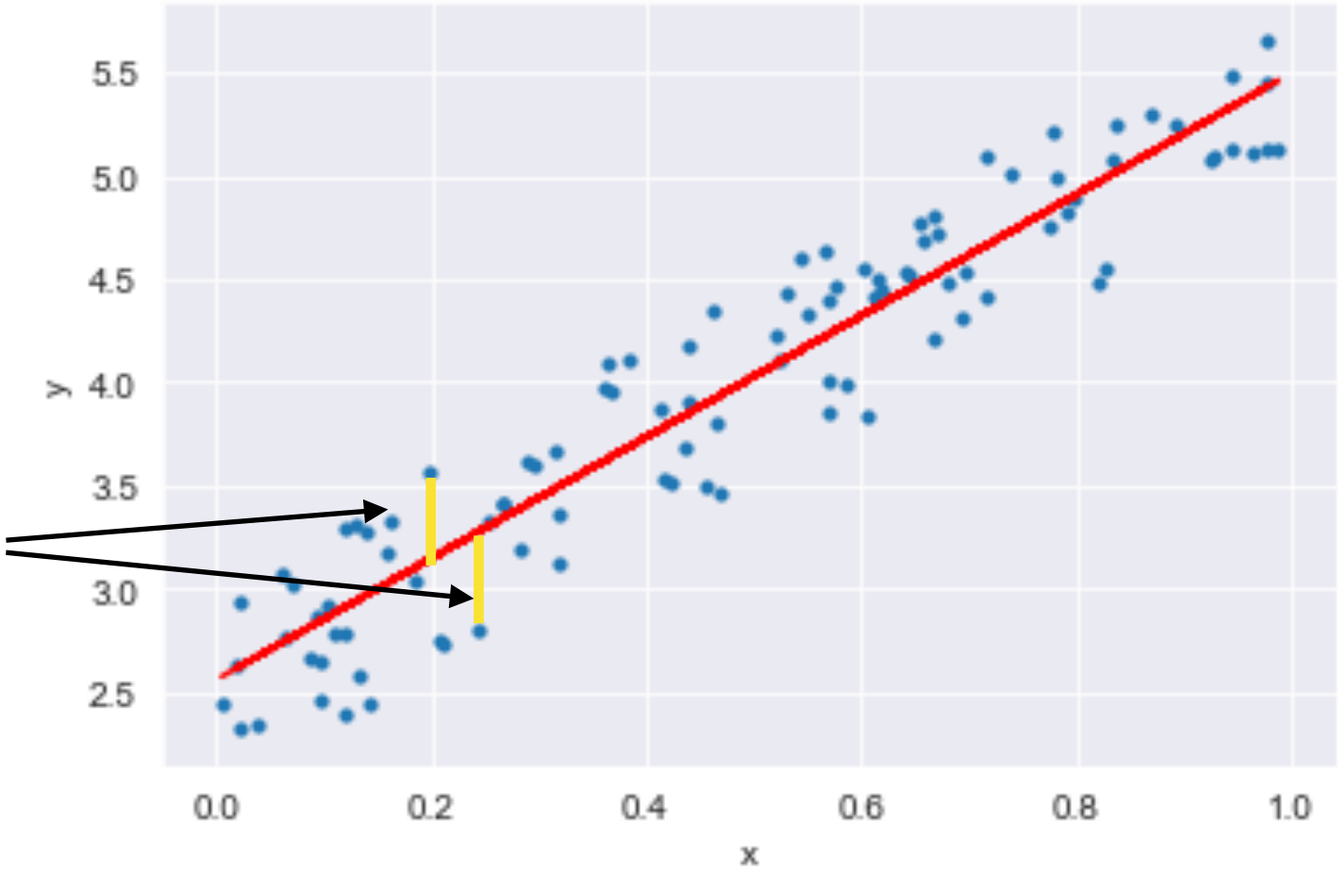
linear regression fits the line

$$y = ax + b$$



Association Mapping

linear regression fits the line

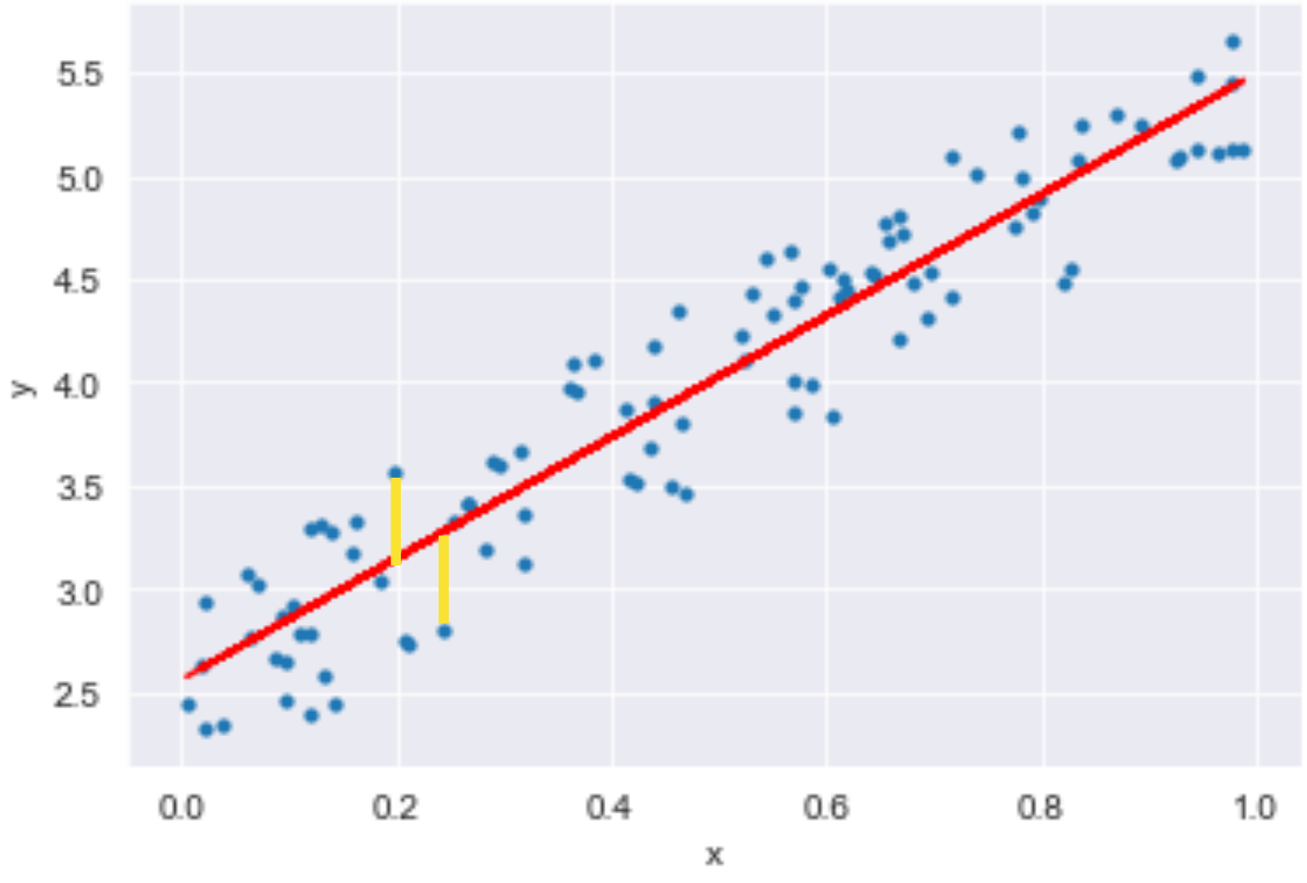


“residuals”

ϵ

Association Mapping

Least Squares method minimizes residuals

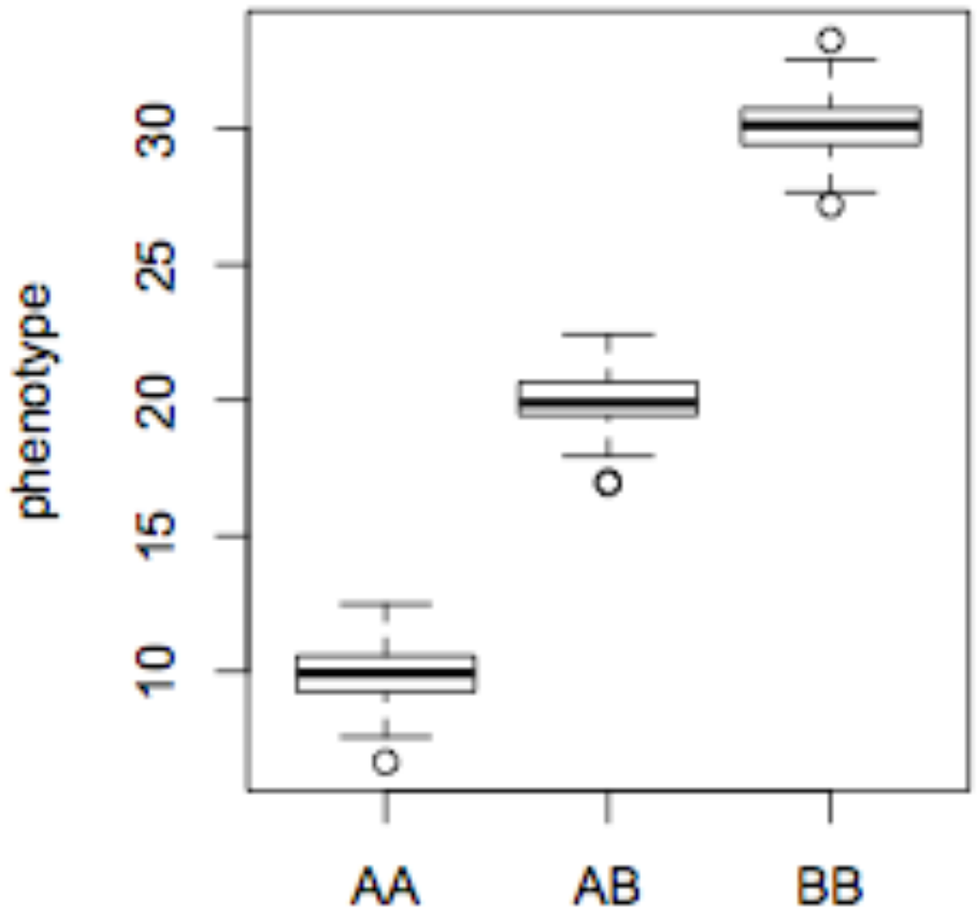


$$\sum_i \epsilon_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

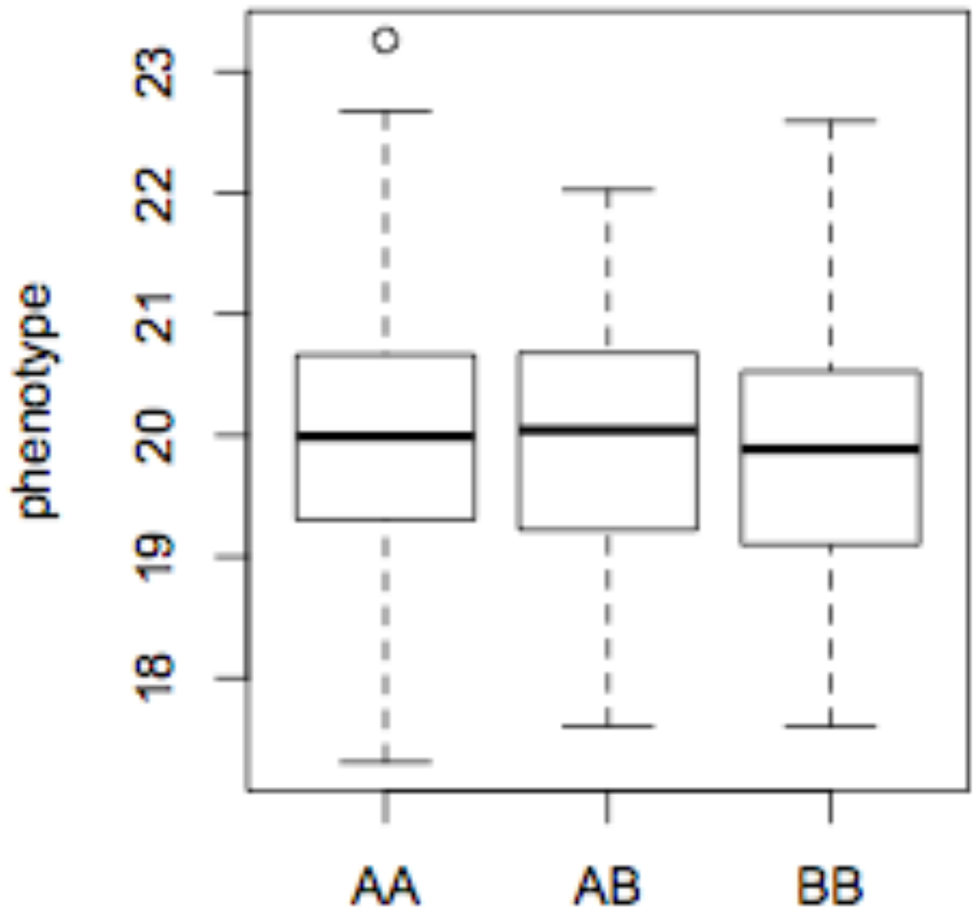
Association Mapping

Use linear regression for this!
pheno = ax + b

Associated Marker



Non Associated Marker



Genome-Wide Association Study

They’re good dogs

Genome-Wide Association Study

Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes

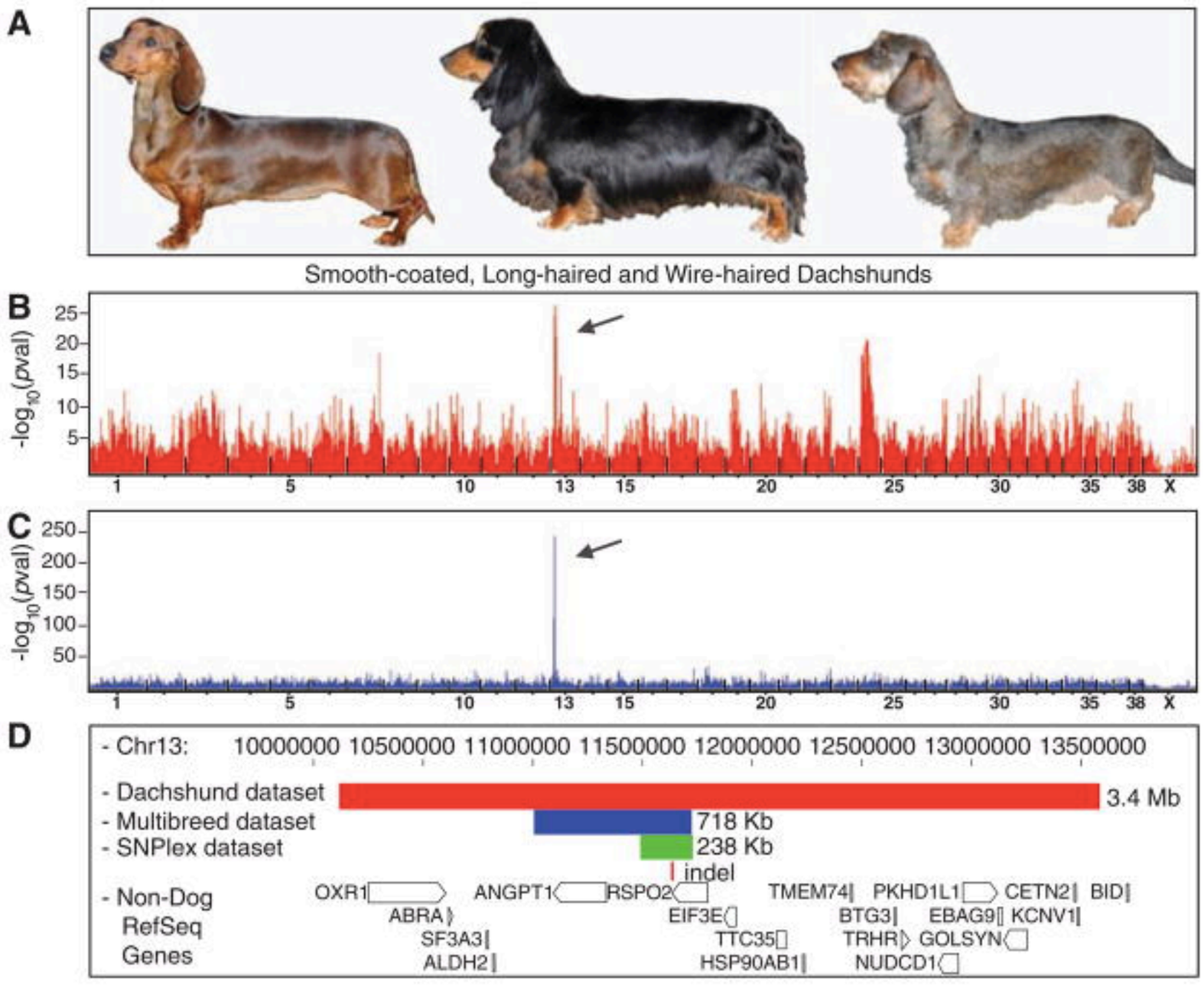
Edouard Cadieu,[1] Mark W. Neff,[2] Pascale Quignon,[1] Kari Walsh,[2] Kevin Chase,[3] Heidi G. Parker,[1] Bridgett M. VonHoldt,[4] Alison Rhue,[2] Adam Boyko,[5] Alexandra Byers,[1] Aaron Wong,[2] Dana S. Mosher,[1] Abdel G. Elkahloun,[1] Tyrone C. Spady,[1] Catherine André,[6] K. Gordon Lark,[3] Michelle Cargill,[7*] Carlos D. Bustamante,[5] Robert K. Wayne,[4] Elaine A. Ostrander[1]†

Coat color and type are essential characteristics of domestic dog breeds. Although the genetic basis of coat color has been well characterized, relatively little is known about the genes influencing coat growth pattern, length, and curl. We performed genome-wide association studies of more than 1000 dogs from 80 domestic breeds to identify genes associated with canine fur phenotypes. Taking advantage of both inter- and intrabreed variability, we identified distinct mutations in three genes, *RSPO2*, *FGF5*, and *KRT71* (encoding R-spondin–2, fibroblast growth factor–5, and keratin-71, respectively), that together account for most coat phenotypes in purebred dogs in the United States. Thus, an array of varied and seemingly complex phenotypes can be reduced to the combinatorial effects of only a few genes.

The tremendous phenotypic diversity of modern dog breeds represents the end point of a >15,000-year experiment in artificial and natural selection (*1*, *2*). As has been demonstrated for traits such as body size (*3*) and coat color (*4*), marker-based associations with phenotypic traits can be explored within single breeds to initially identify regions of genetic association, and then expanded to multiple breeds for fine-mapping and mutation scanning (*5*, *6*).

Genome-Wide Association Study

GWAS of dog mustache



Genome-Wide Association Study

GWAS of curly hair

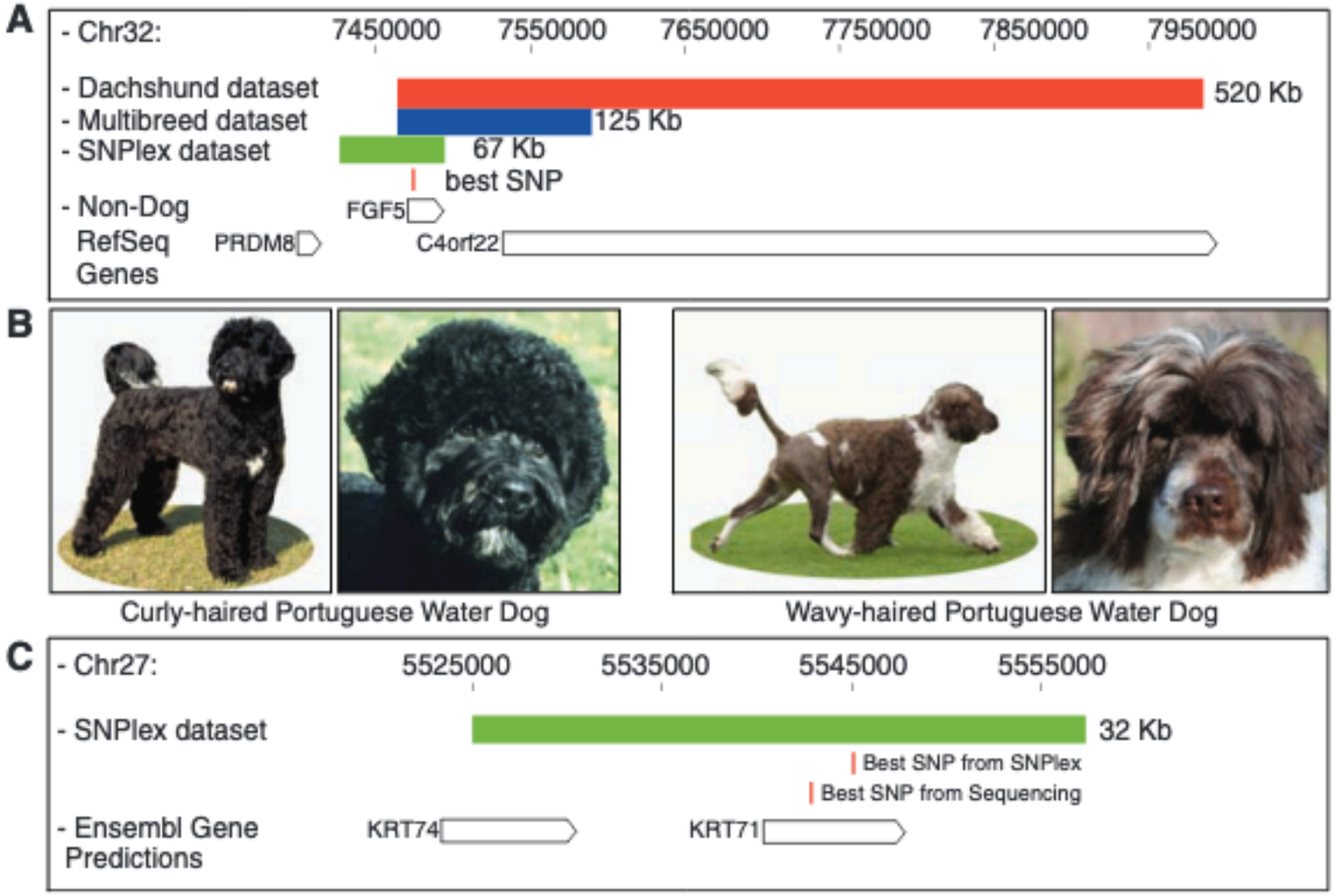


Fig. 2. Regions of homozygosity identify genes for pelage length and curl. (**A**) Homozygous region found on CFA32 defining the length locus. The red bar indicates the 520-kb associated haplotype from 29 long-haired dachshunds; the blue bar spans the 125-kb homozygous region found in 319 dogs from 31 long-haired breeds; the green bar represents the 67-kb reduced homozygous region found after fine-mapping in 293 dogs from 39 long-haired breeds. The best-associated SNP, represented by the small red rectangle, is within these three homozygous regions in exon 1 of the

Genome-Wide Association Study

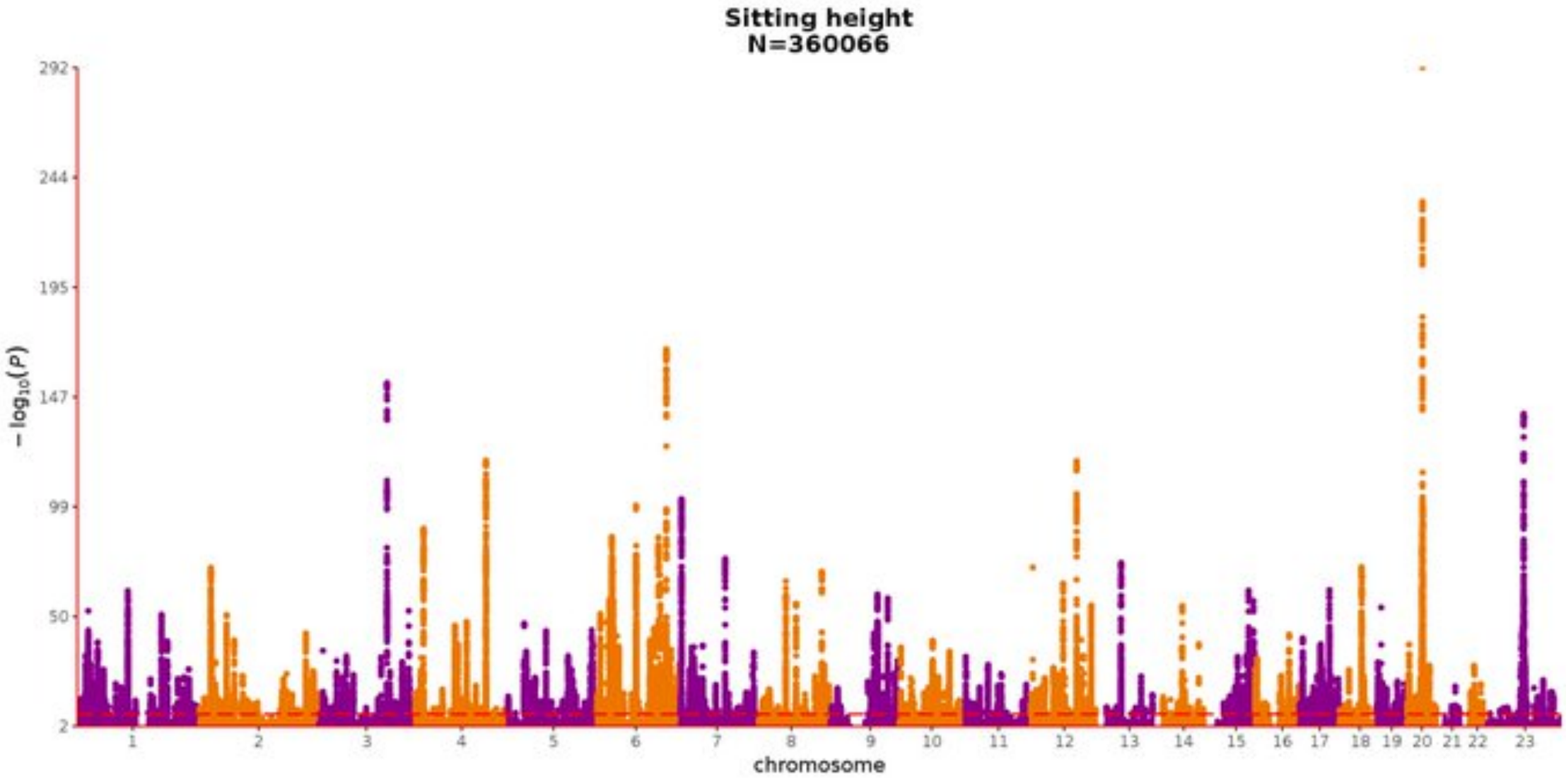
large effect alleles explain a lot of variation

	PHENOTYPE	*FGF5*	*RSPO2*	*KRT71*
A	Short	-	-	-
B	Wire	-	+	-
C	Wire and Curly	-	+	+
D	Long	+	-	-
E	Long with Furnishings	+	+	-
F	Curly	+	-	+
G	Curly with Furnishings	+	+	+

Fig. 3. Combinations of alleles at three genes create seven different coat phenotypes. Plus (+) and minus signs (–) indicate the presence or absence of variant (nonancestral) genotype. A characteristic breed is represented for each of the seven combinations observed in our data set: (**A**) short hair; (**B**) wire hair; (**C**) "curly-wire" hair; (**D**) long hair; (**E**) long, soft hair with furnishings; (**F**) long, curly hair; and (**G**) long, curly hair with furnishings. [Photos courtesy of M. Bloom (Copyright AKC)].

Genome-Wide Association Study

Human variation doesn’t look like this!



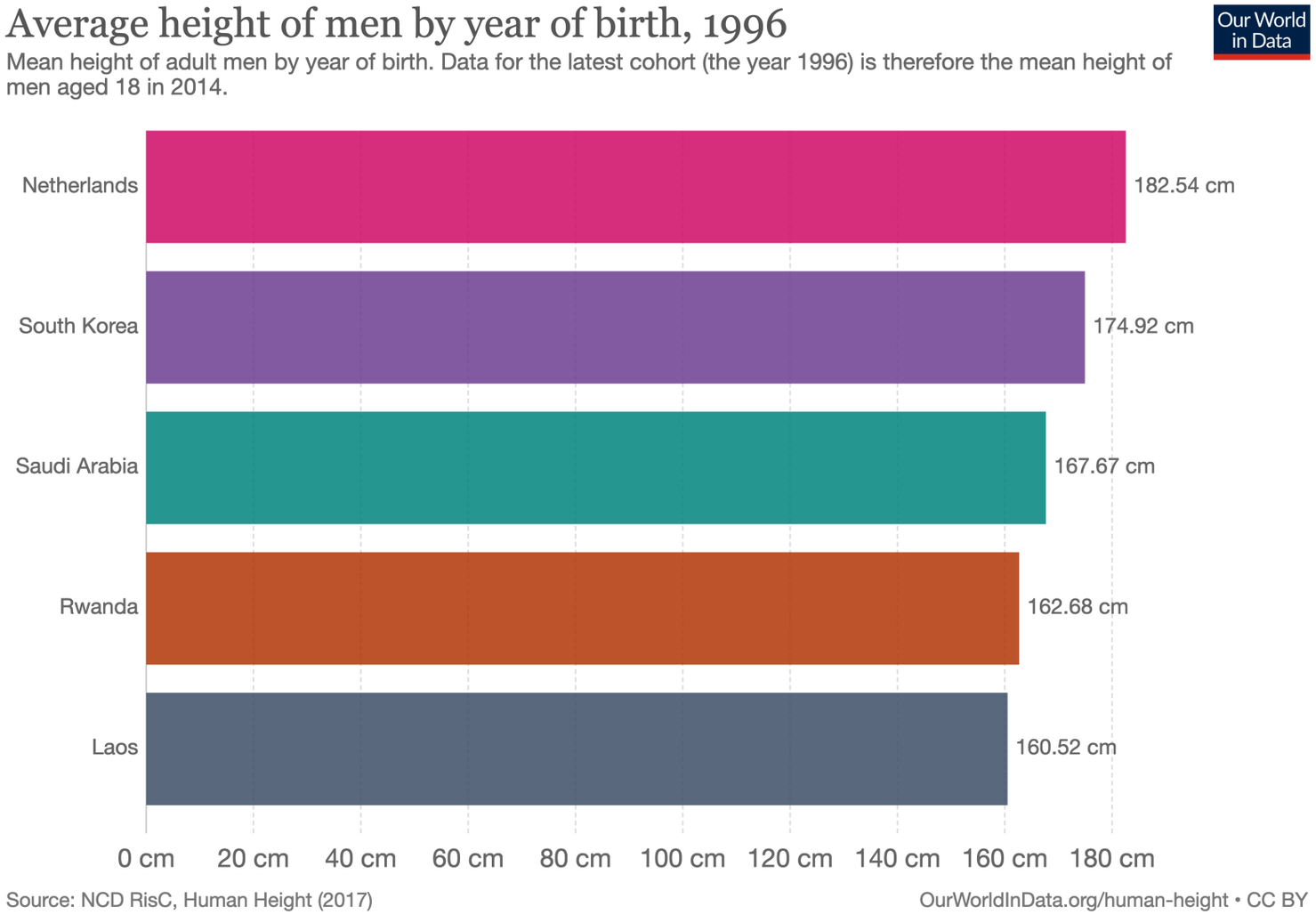
Sitting Height @SbotGwa

Issues with GWAS

it’s a long list

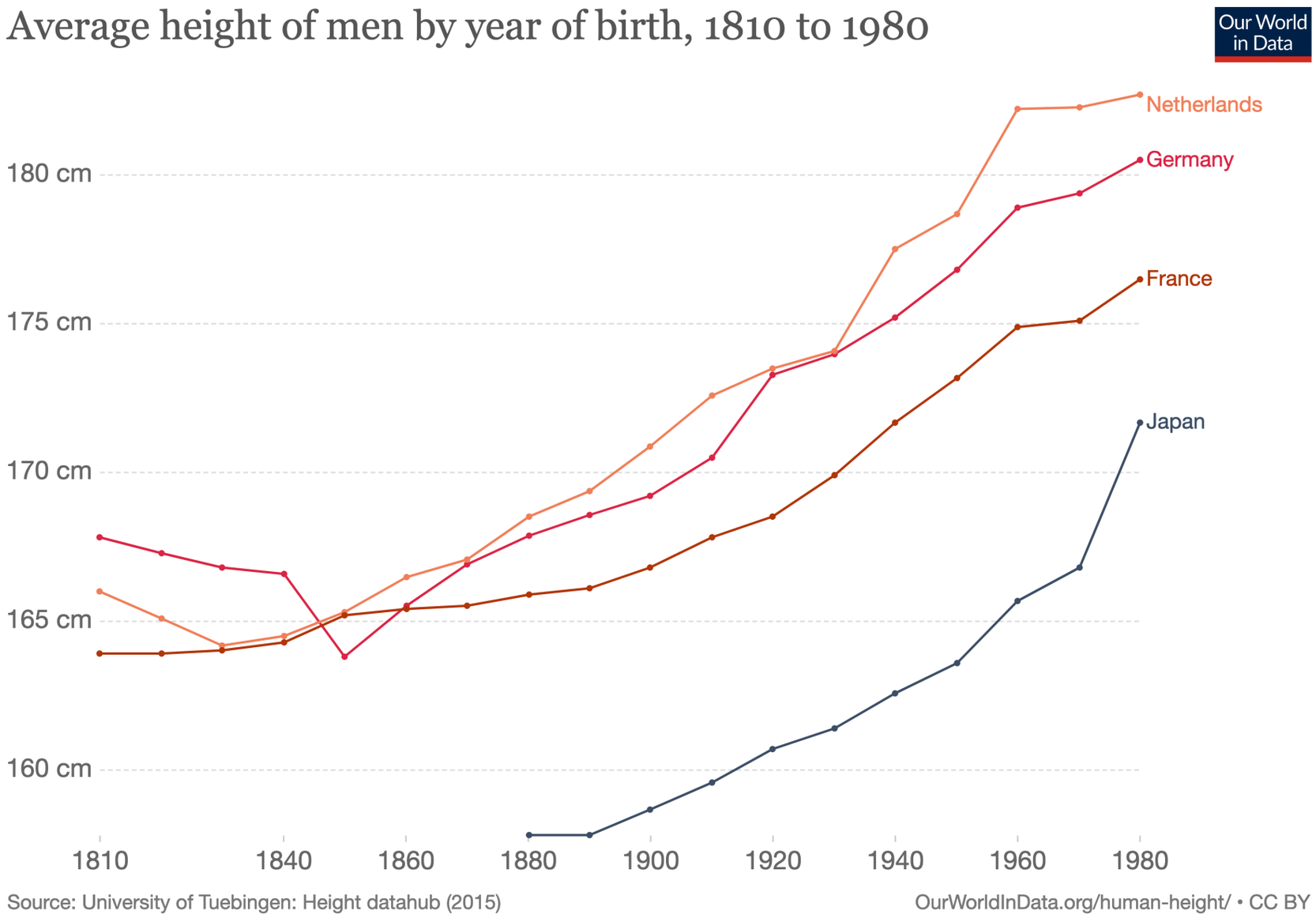
- Interpretation- what is heritability?
- Population Structure
- Biased estimates of β
    - “Winner’s curse”

Interpretation



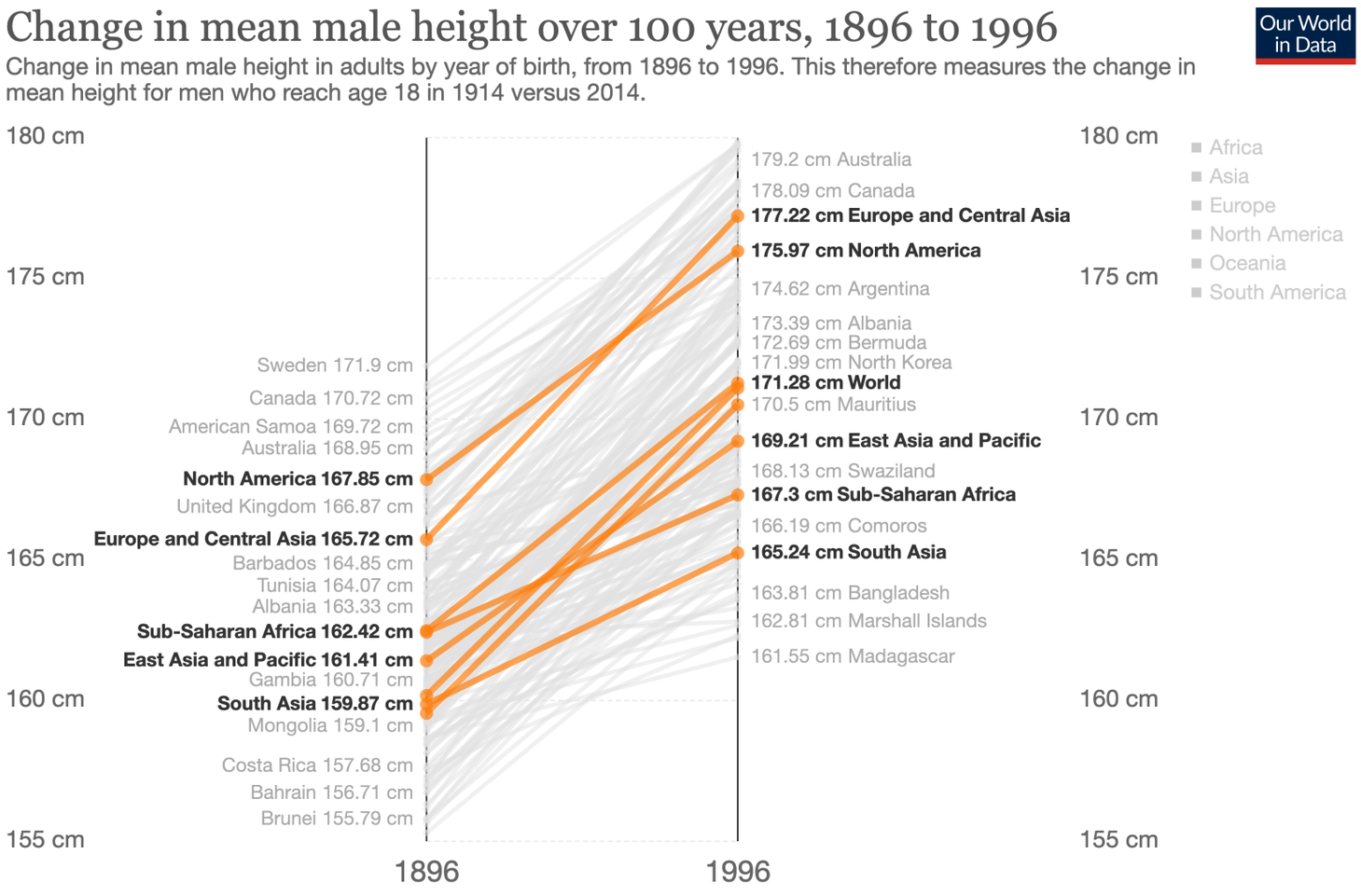
$$V_p \sim V_g + V_e$$

Interpretation



but phenotypes change with environment!

Interpretation



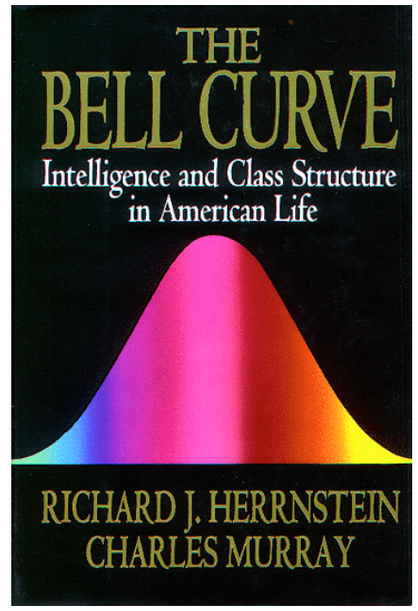
Also *perhaps* GxE components

“The Bell Curve”

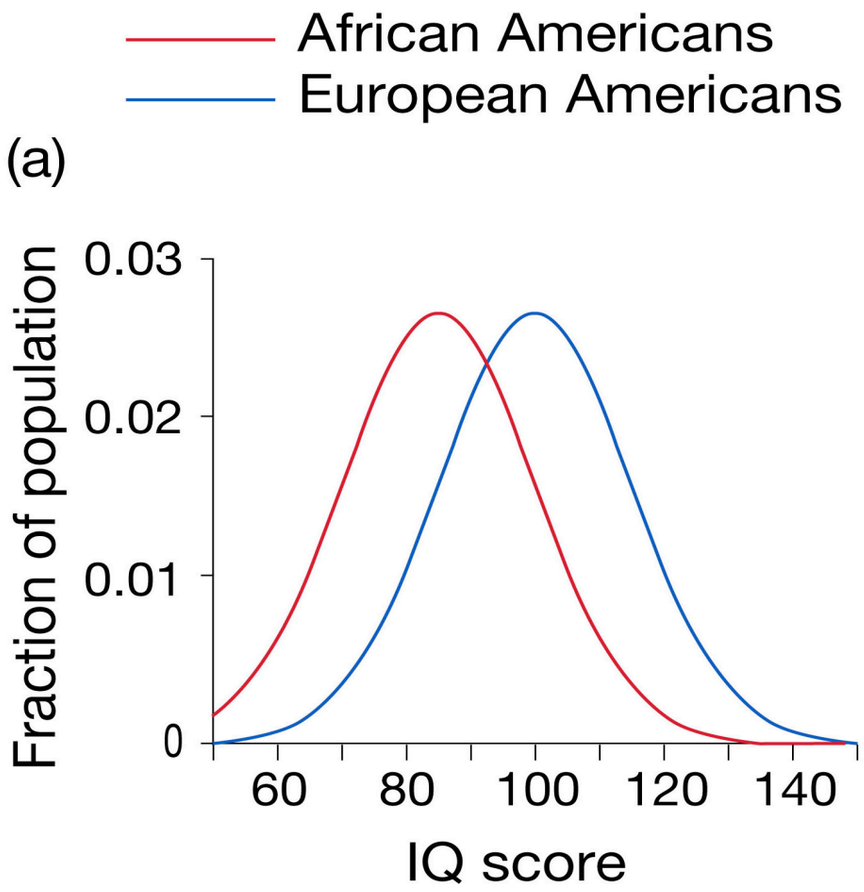
misinterpreting quantitative genetics 101

Argument:

1) Heritable variation for IQ, i.e. $\frac{V_g}{V_p} = h^2 > 0$

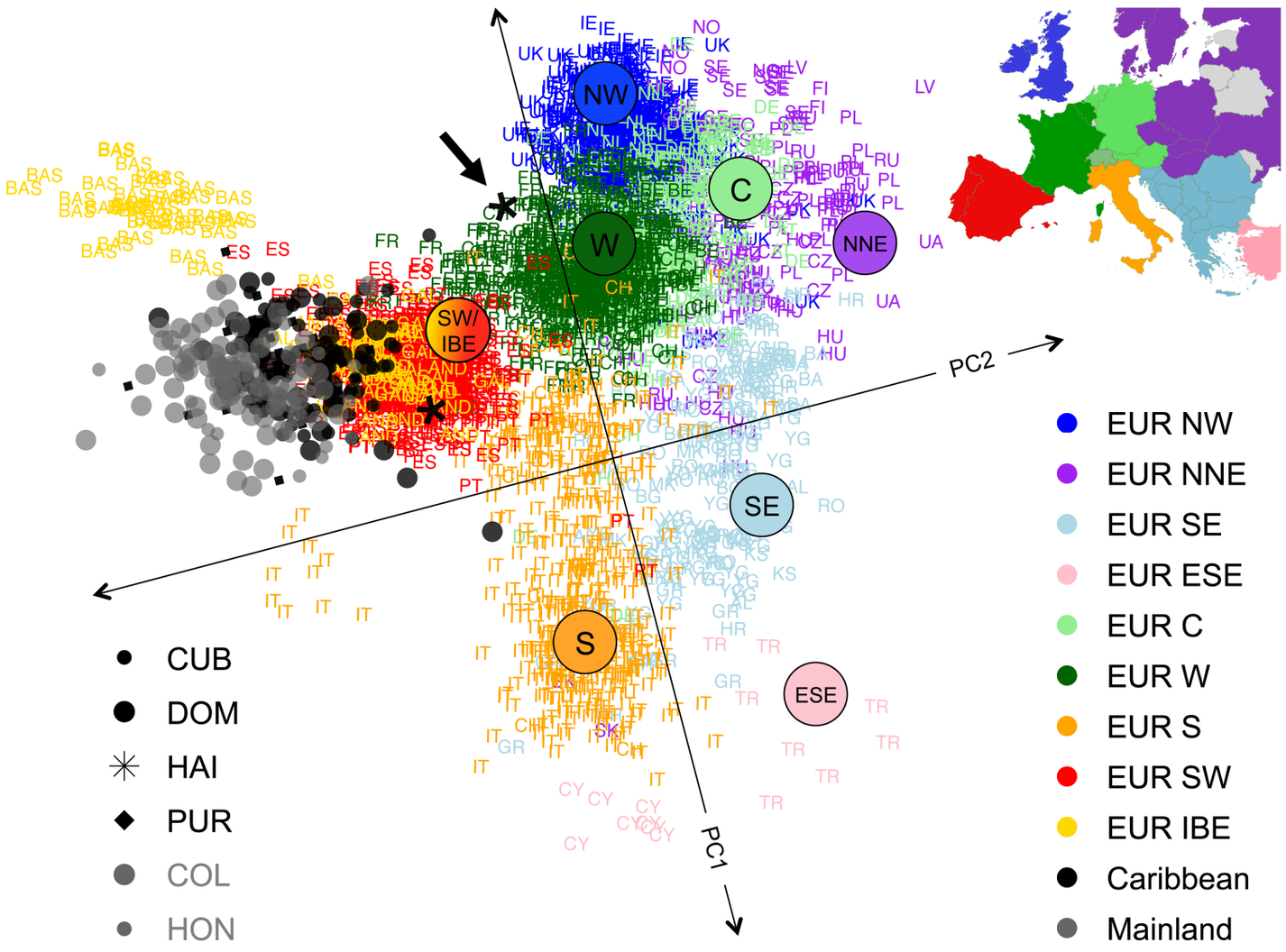


2) Phenotypic means are different between populations



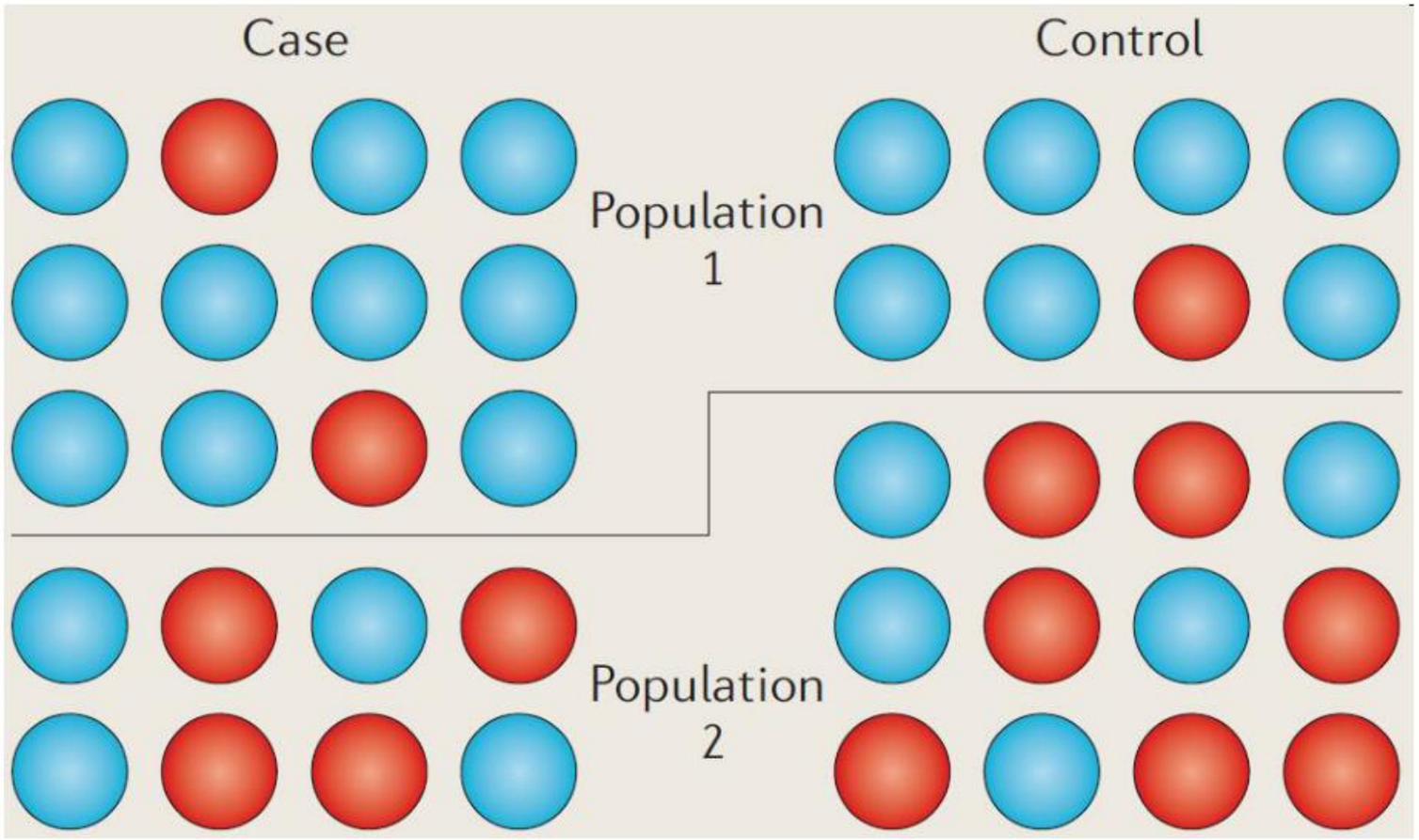
So what’s the problem here?

Population structure



genetic differences in allele frequencies at unassociated loci!

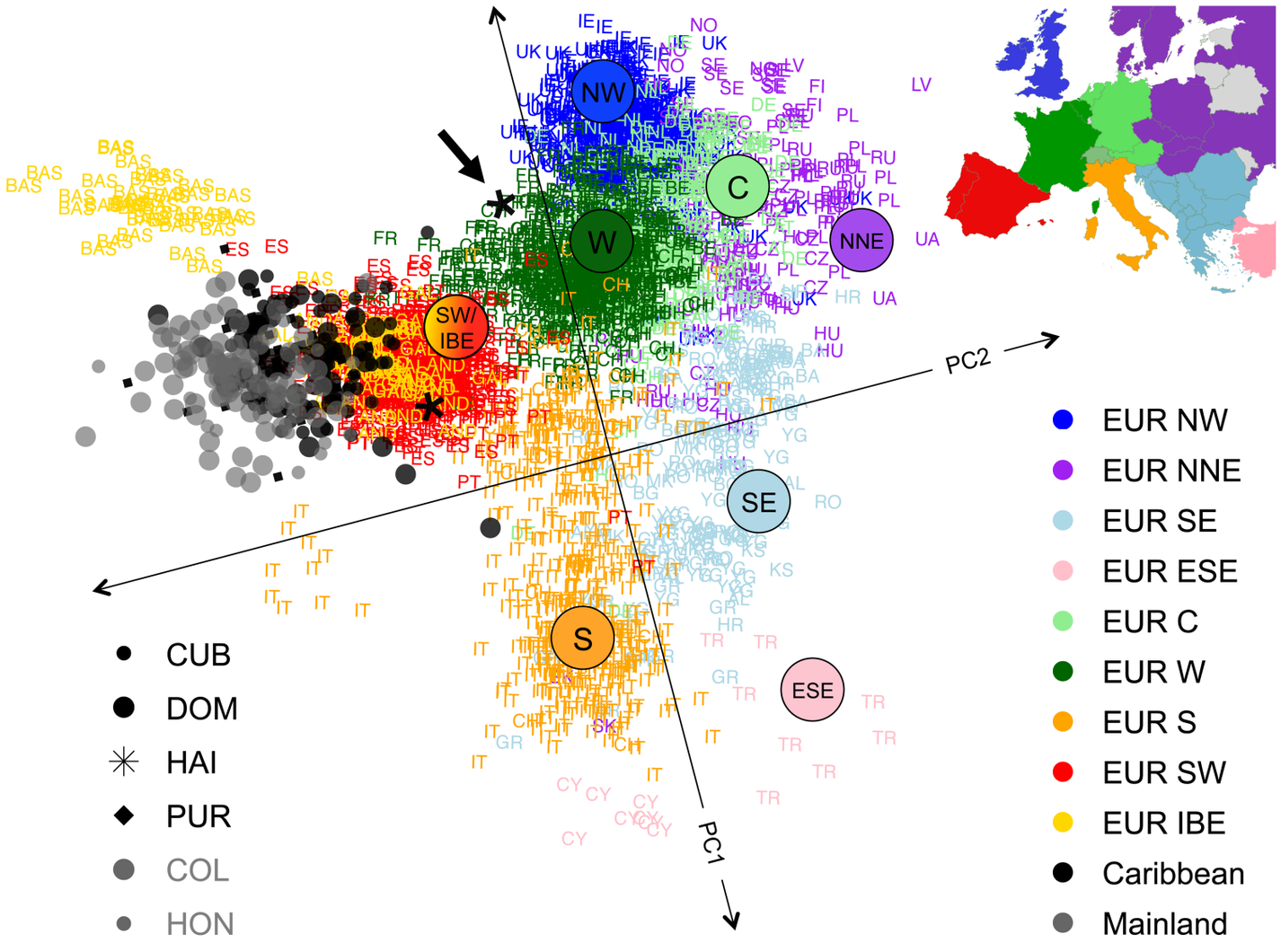
Population structure



Balding, Nature Reviews Genetics 2010

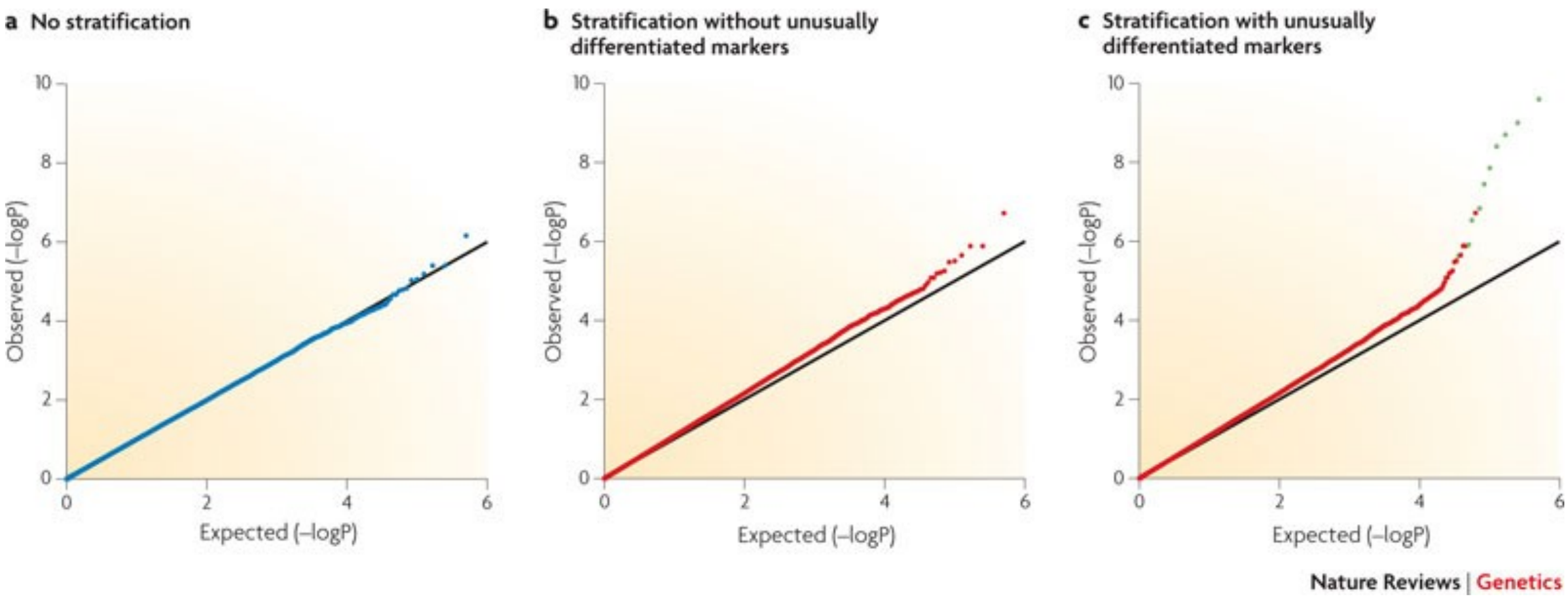
genetic differences in allele frequencies at unassociated loci!

Population structure



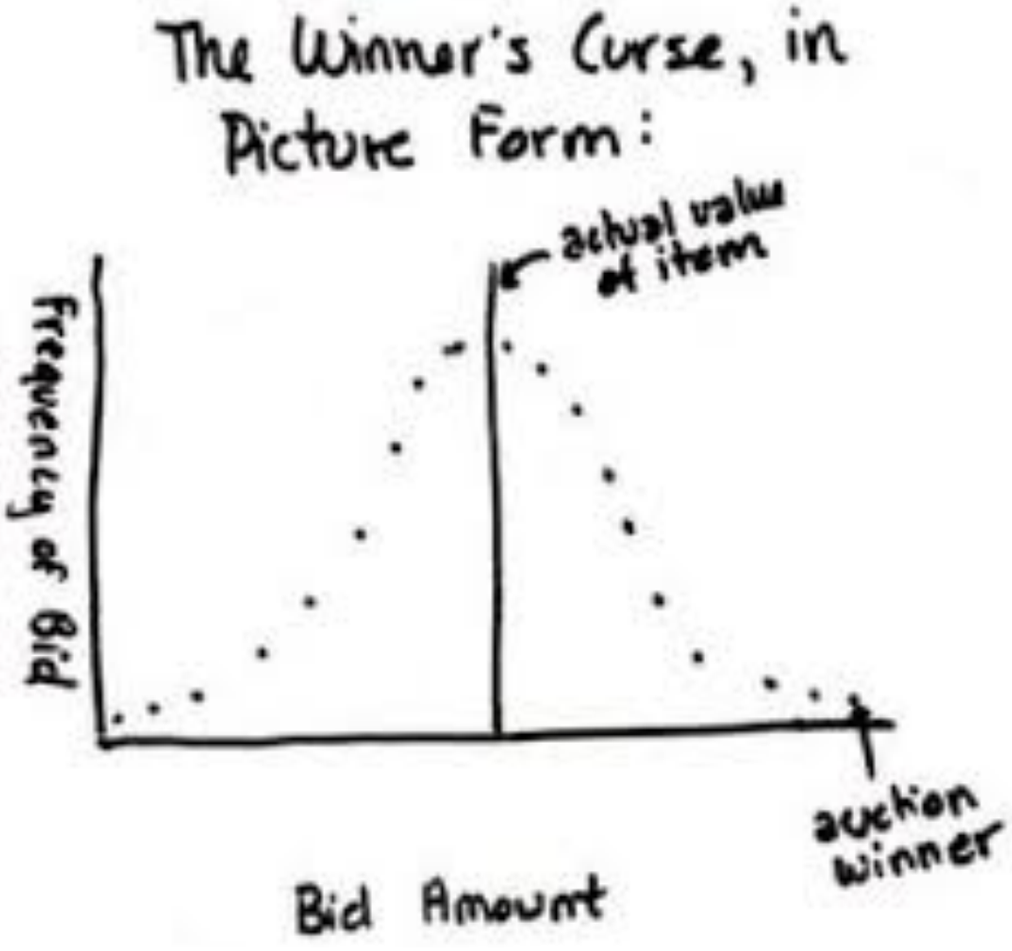
will correct for this directly in our linear model!

Population structure



will correct for this directly in our linear model!

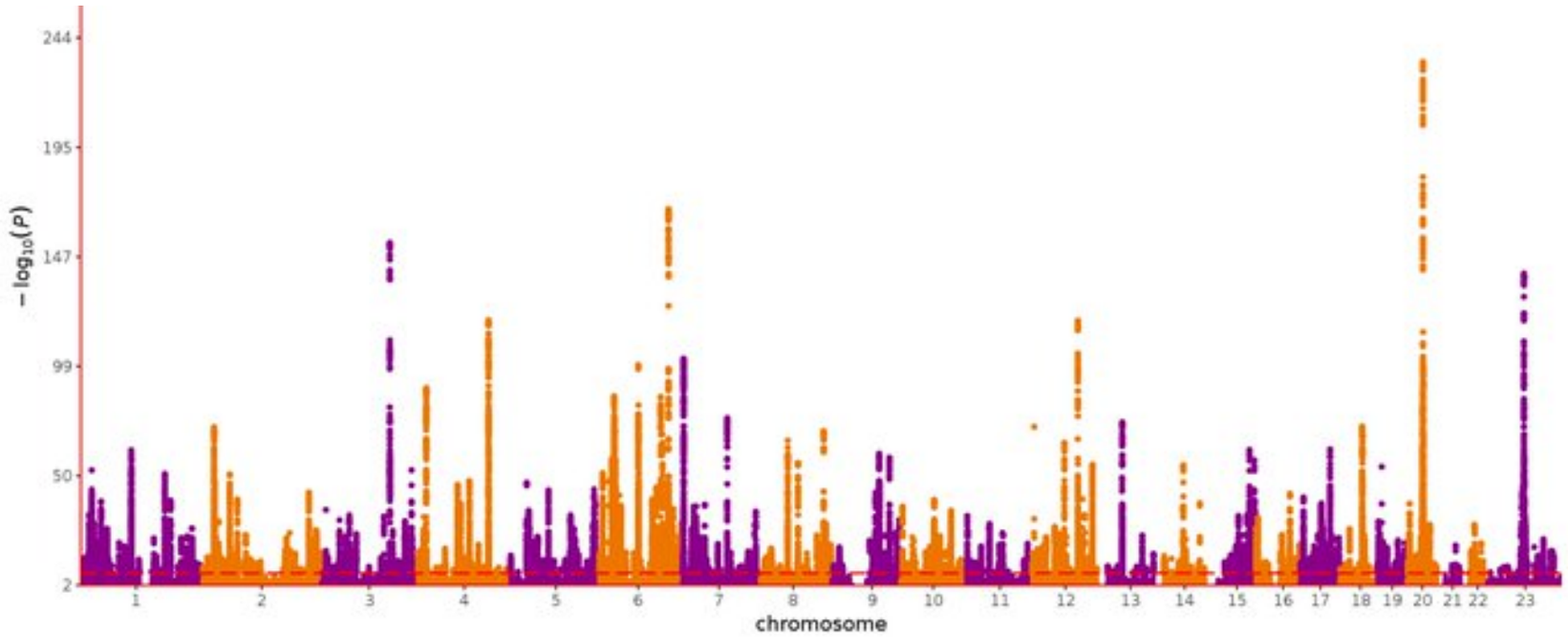
Winner's Curse



In GWAS we should overestimate βs in underpowered studies

Polygenic scores

effect size matters!



Easy way to do phenotypic prediction would be to sum up effect sizes!

$$PGS = \sum_{i}^{L} X_i \beta_i$$

effect size estimate at locus i

genotype at locus i

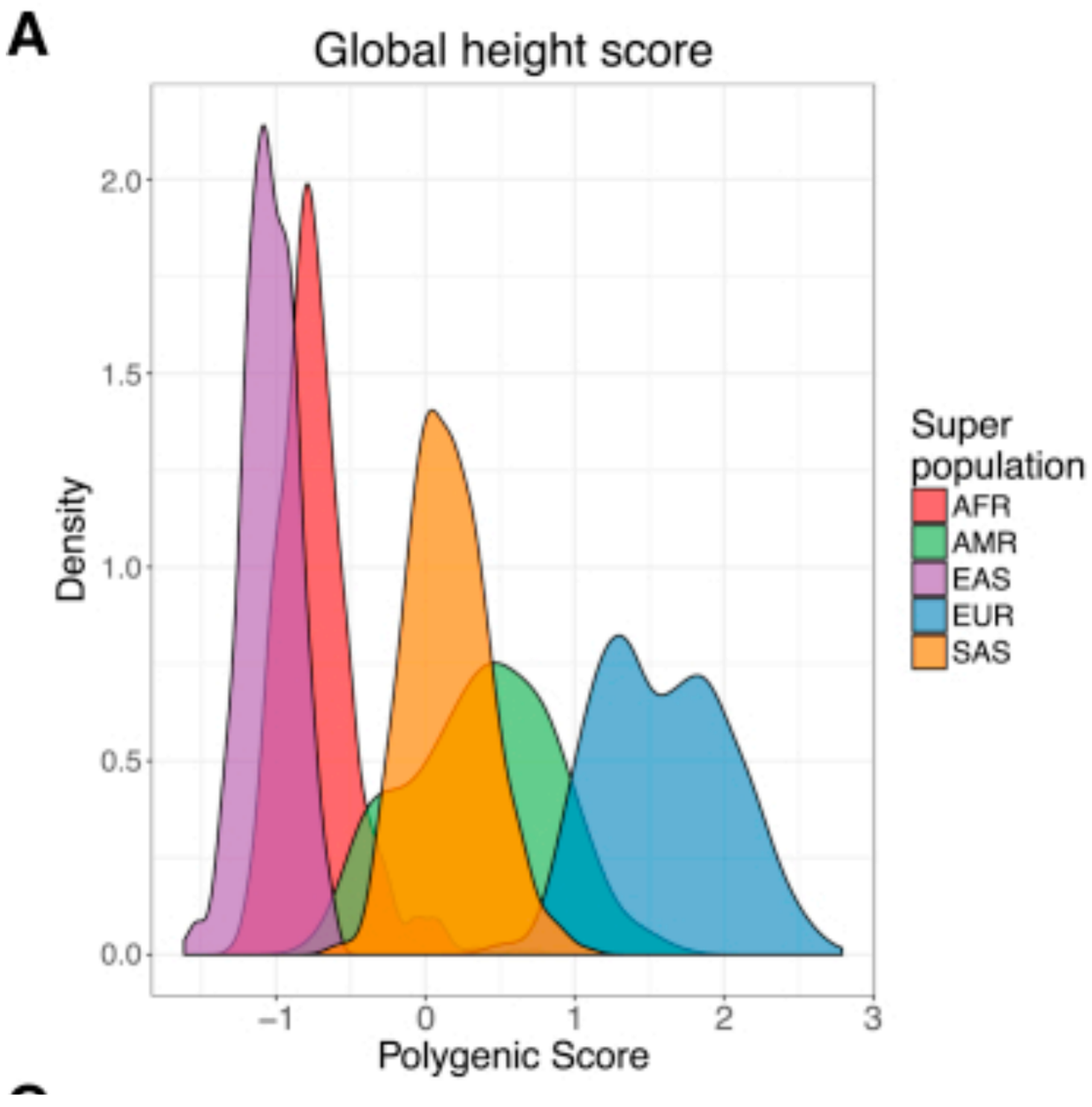
Polygenic scores

$$PGS = \sum_{i}^{L} X_i \beta_i$$

So what are the problems with this?

Where are the effect size estimates coming from?

Polygenic scores



One huge issue is effect size estimates come from GWAS!