Pop Gen meets Quant Gen and other open questions

Ryan D. Hernandez

qb3 ucb·ucsc·ucsf

UCSF University of California San Francisco

Department of Bioengineering and Therapeutic Sciences
a joint department of the UCSF Schools of Pharmacy and Medicine

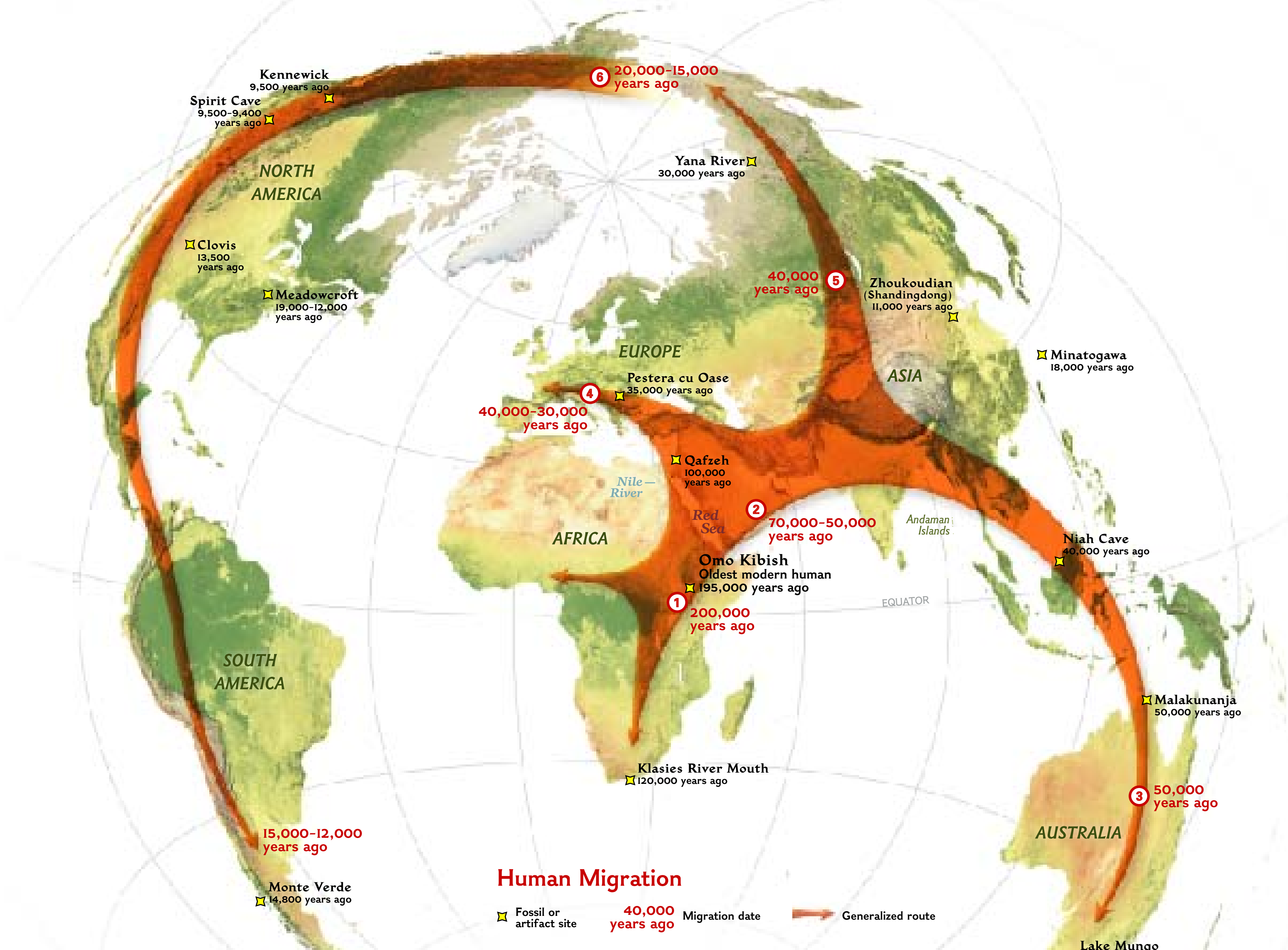
McGill

ryan.hernandez@me.com

Modern Human Genomics

http://www.finca.org

Human Colonization of the World



SOURCES: SUSAN ANTÓN, NEW YORK UNIVERSITY; ALISON BROOKS, GEORGE WASHINGTON UNIVERSITY; PETER FORSTER, UNIVERSITY OF CAMBRIDGE; JAMES F. O'CONNELL, UNIVERSITY OF UTAH; STEPHEN OPPENHEIMER, OXFORD UNIVERSITY; SPENCER WELLS, NATIONAL GEOGRAPHIC SOCIETY; OFER BAR-YOSEF, HARVARD UNIVERSITY

NGM MAPS

http://ngm.nationalgeographic.com

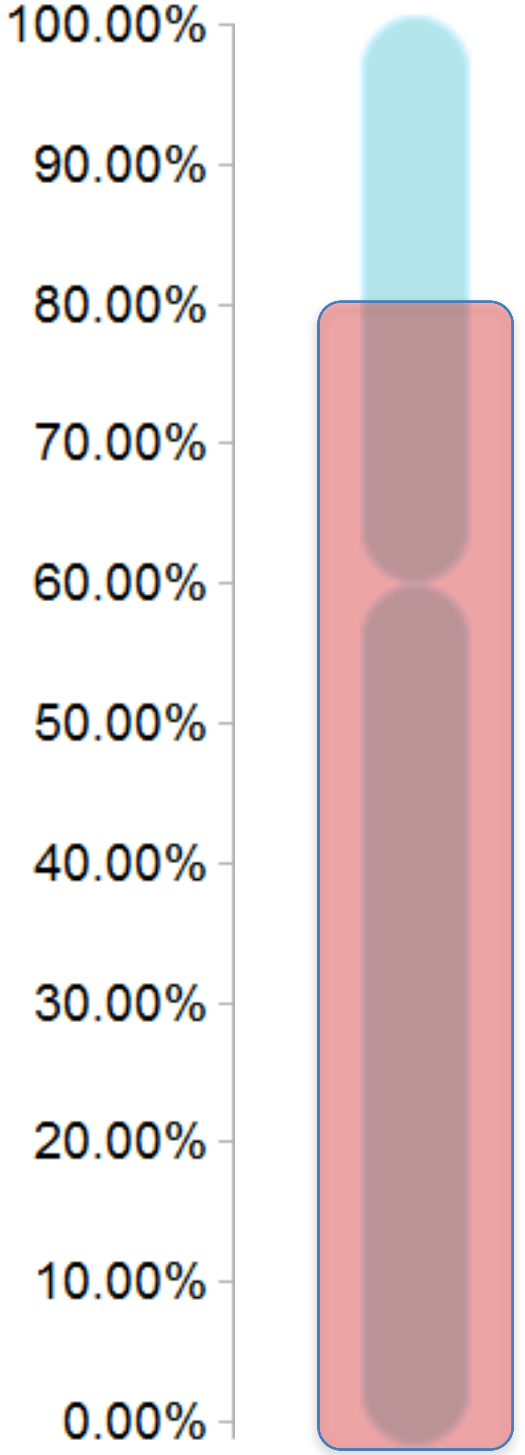
Heritability and Human Height

Studies of heritability ask questions such as how much genetic factors play a role in **differences in height between people**. This is not the same as asking **how much genetic factors influence height in any one person.**

https://en.wikipedia.org/wiki/Heritability

http://i.ytimg.com/vi/E0Aeks_id6c/maxresdefault.jpg

An estimated 80% of variation in height driven is driven by genetics

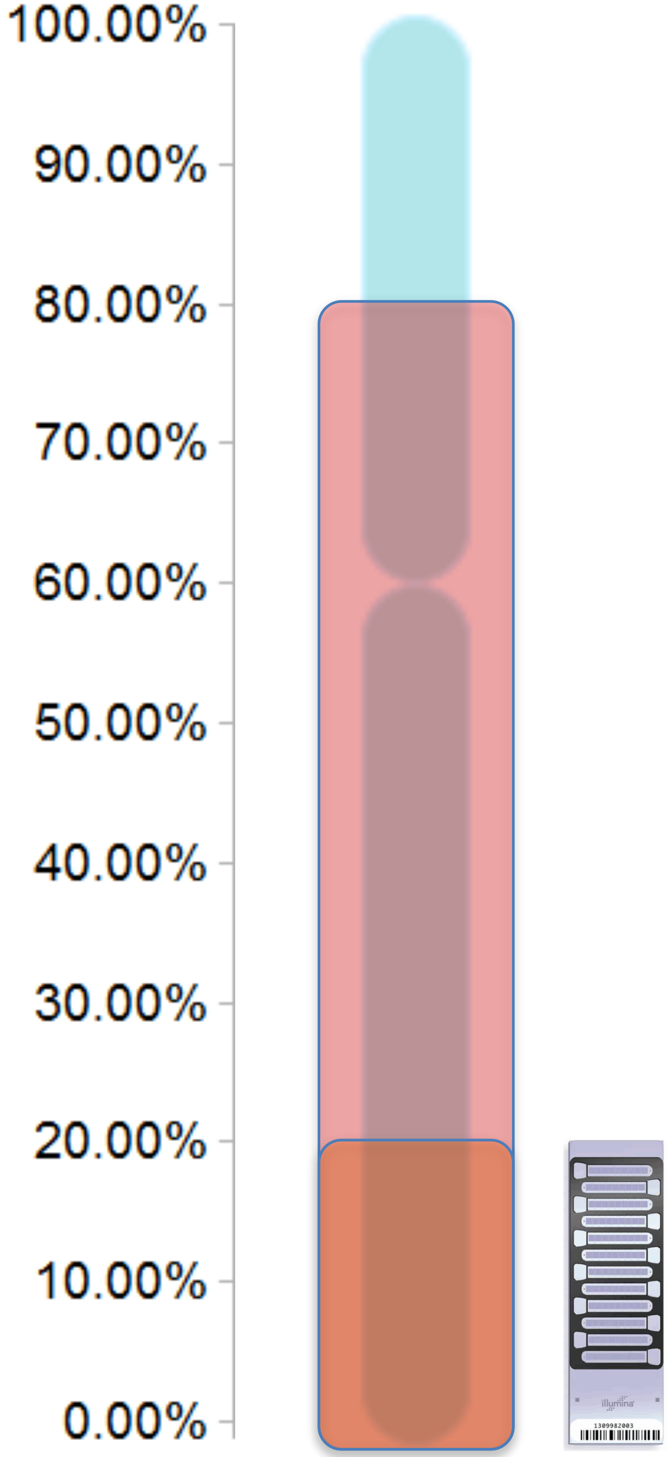


Large twin study

Silventoinen et al, 2003 Twin Research

http://i.ytimg.com/vi/E0Aeks_id6c/maxresdefault.jpg

But GWAS explain only 20% of the variation in height

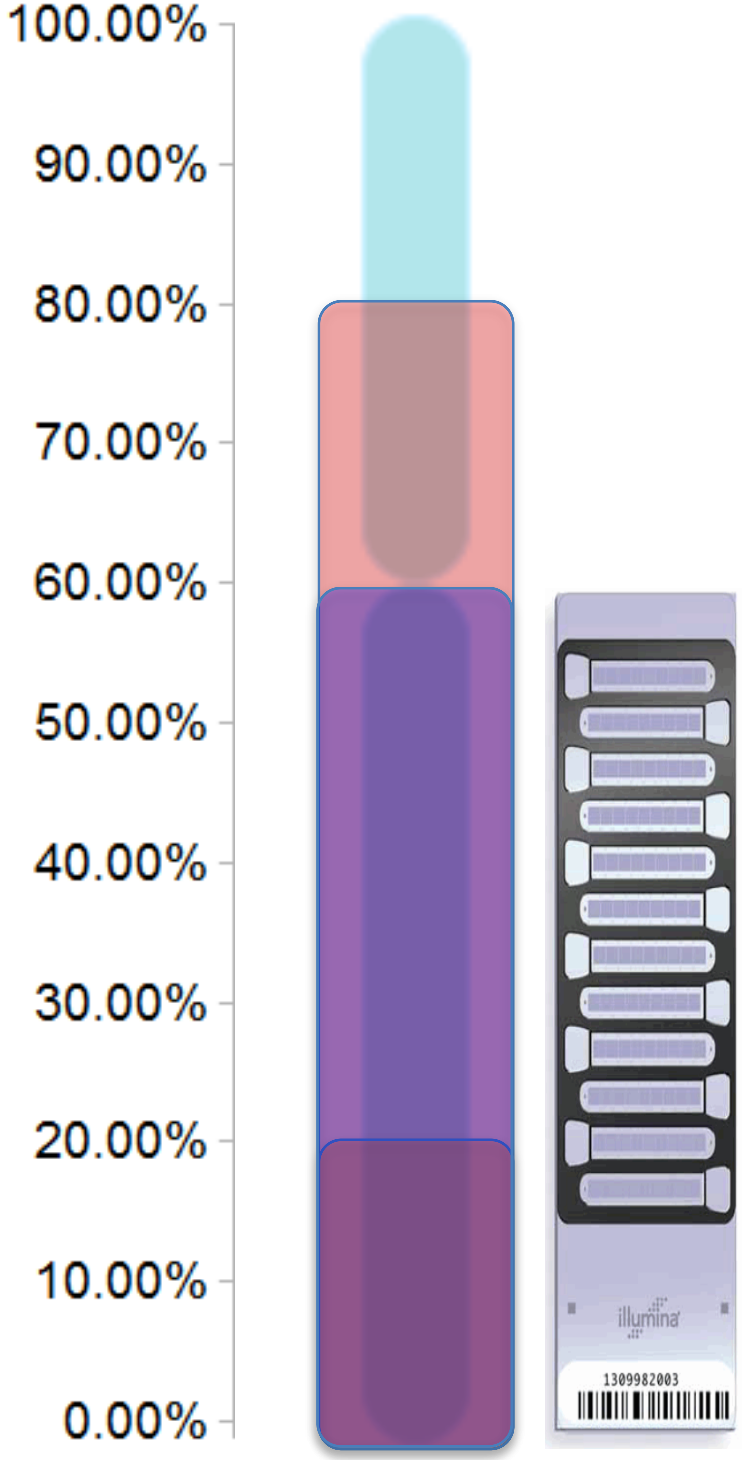


h^2_{GWAS} : The narrow-sense heritability explained by summing the effects of GWAS identified SNPs.

250,000 subjects

Wood et al, 2014 Nat. Genet.
i.ytimg.com/vi/E0Aeks_id6c/maxresdefault.jpg

GWAS have the potential to explain 60% of the variation in height



h_g^2 : The narrow-sense heritability explained by all genotyped SNPs.

250,000 subjects

Wood et al, 2014 Nat. Genet.
i.ytimg.com/vi/E0Aeks_id6c/maxresdefault.jpg

Challenges For Studying Complex

The case of the missing heritability

Maher, *Nature* (2008).

Major Problem

- There are no complex traits in which we know:
  - The number of causal variants
  - The frequencies of all the causal variants
  - The effect sizes of all the causal variants
  - The fitness effect of all the causal variants
- We need a thorough simulation study where we can vary all of these parameters and see how they effect our answer!

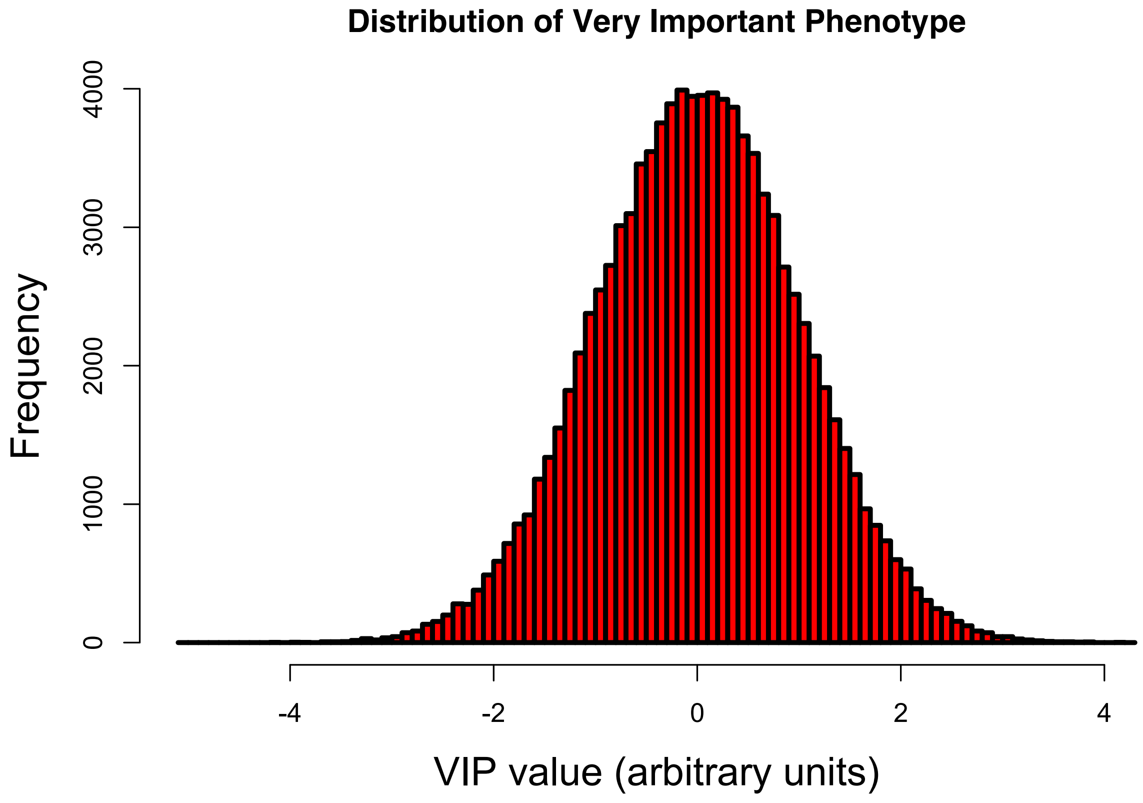
Possible Origins Of Missing Heritability

Candidates
Common variants of weak effect
Incomplete linkage to causal alleles/multiple causal alleles in locus
GxG / GxE Interactions
Rare variants
Structural variation

From GWAS To Deep Sequencing

- Genome-wide association studies (GWAS) seek to identify common variants that contribute to common disease
- Successfully identified many candidate disease-associated genes
- Challenges:
    - Generally have low relative risk
    - Explain only a small proportion of the phenotypic variance
    - Provides candidate loci, but causal variant is rarely typed
- Implication:
    - Predictive power of GWAS is minimal…

“Missing” heritability - calculating variance accounted for by GWAS



Suppose *k* variants are found to be associated with VIP…

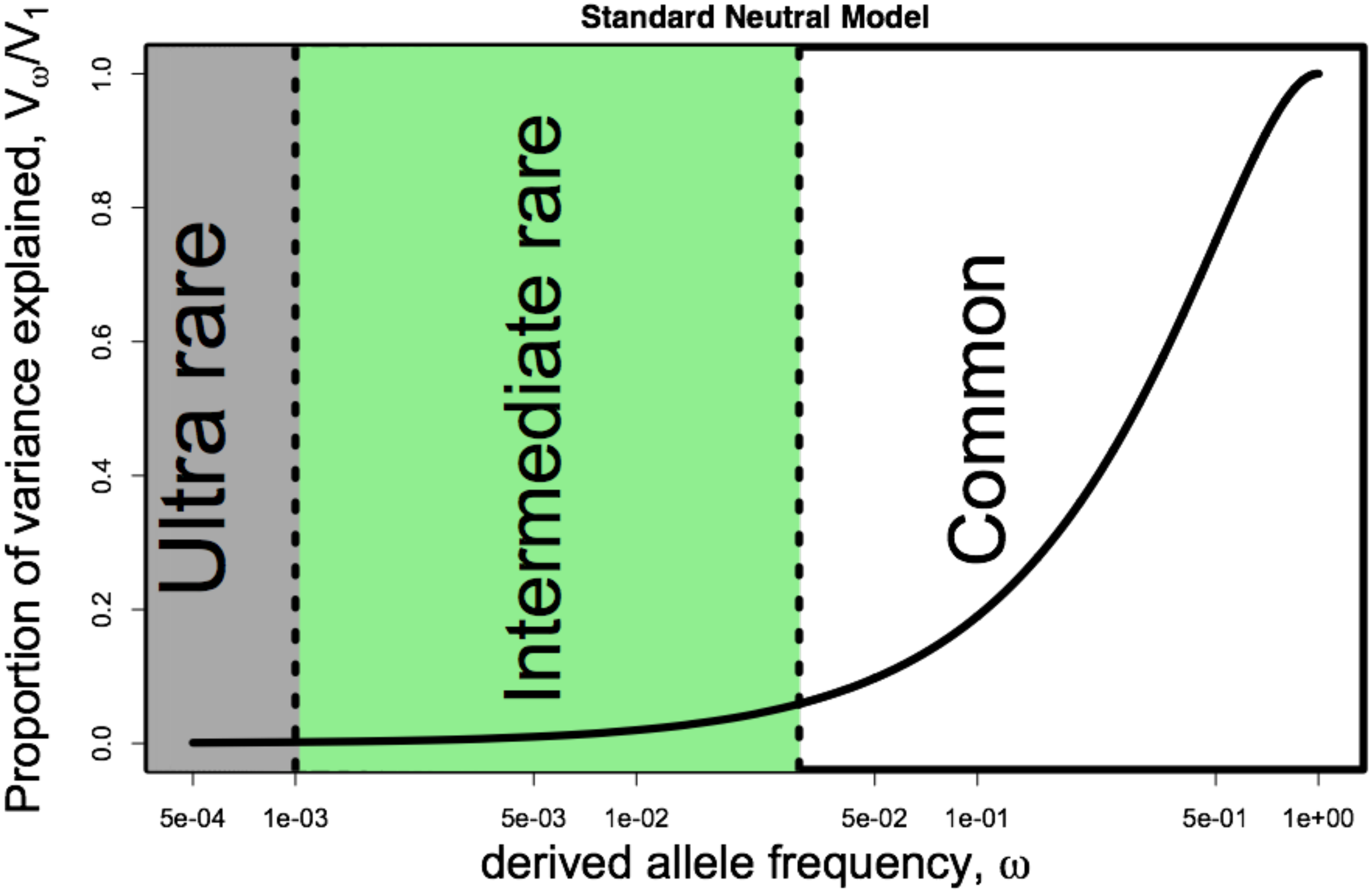
Contribution from each SNP $$v = \frac{1}{2} z^2 x(1-x)$$

Total variance from GWAS $$V_{\text{GWAS}}(P) = \sum_k v_k$$

Compare to GWAS $$V_{\text{GWAS}}(P) < h^2 \times V(P)$$

Lawrence Uricchio

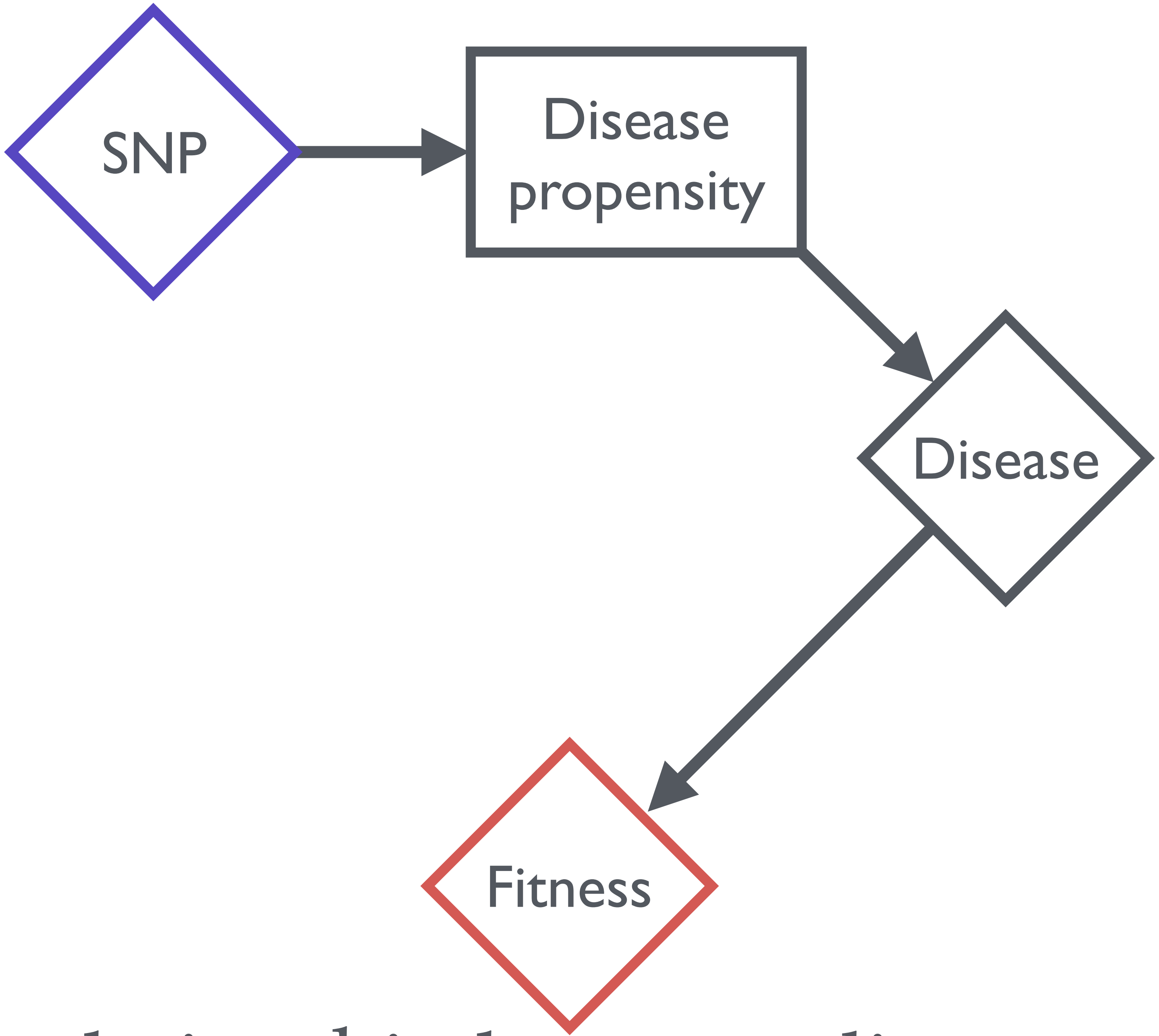
Where is the "missing" heritability?



Population Genetics

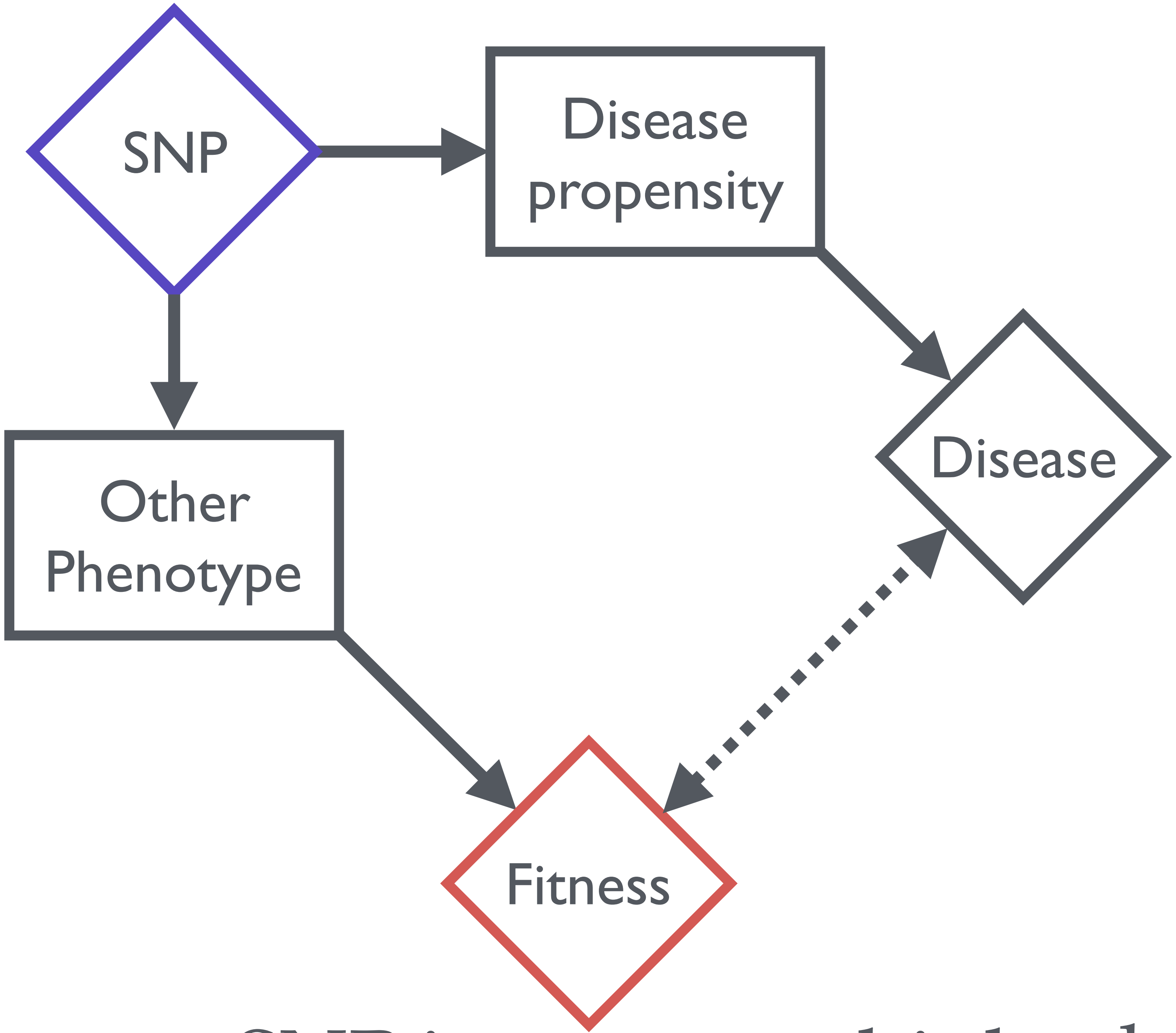
- Why would cases have an excess of **rare** non-synonymous variants in disease-associated genes?
  - Recent neutral mutations that have not had time to spread
  - Deleterious mutations restricted to low frequency
- Population genetic analyses are ideally suited to distinguish these cases.

Evolutionary Models Of Complex Disease



Direct relationship between disease and fitness

Evolutionary Models Of Complex Disease



Pleiotropy: SNP impacts multiple phenotypes

Uricchio et al., Genome Research (2016)

The Model Of Eyre-Walker (2010)

- The phenotypic effect size has a direct relationship to selection coefficient of causal mutations:

$$z = \delta S^{\tau}(1 + \epsilon)$$

- Where:
  - $\varepsilon \sim N(0, \sigma^2)$
  - δ = random sign (trait increasing / decreasing)
  - S = selection coefficient
  - τ = measures how the mean absolute effect of a mutation on the trait increases with the strength of selection

Eyre-Walker, PNAS (2010)

The Model Of Simons Et Al (2014)

- The phenotypic effect size **may** have a direct relationship to selection coefficient of causal mutations:

$$z_s \propto \begin{cases} s & \text{with probability } \rho \\ s_r & \text{with probability } (1-\rho) \end{cases}$$

- Where:
  - ρ = Probability that the trait effect is proportional to the selection coefficient: **Pleiotropy**!!
  - s = selection coefficient
  - s_r = random selection coefficient

Simons et al, Nat Genet (2014)

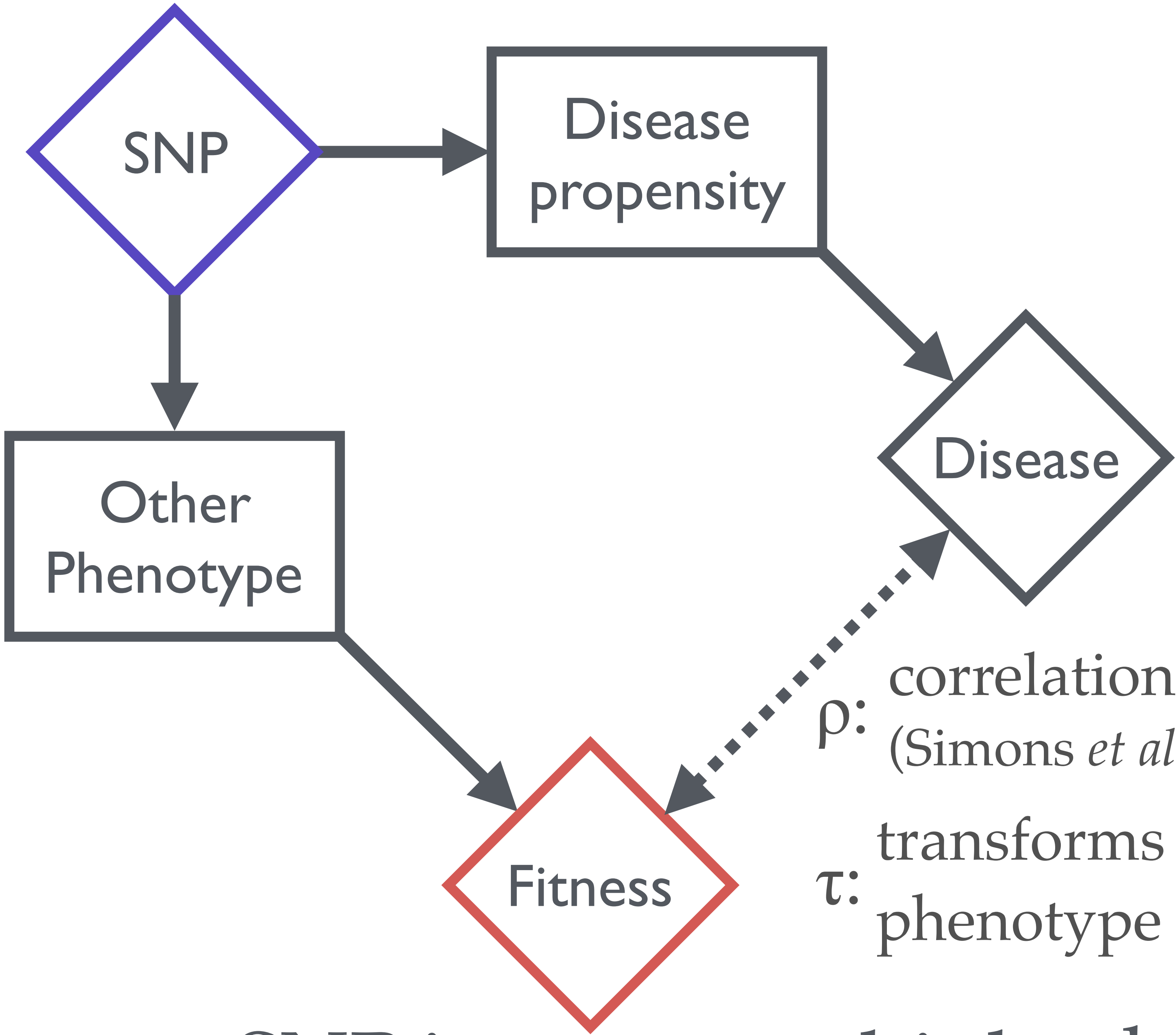
The Model Of Uricchio Et Al (2016)

- A hybrid of the two: $z_s \propto \begin{cases} \delta|s|^{\tau} & \text{with probability } \rho \\ \delta|s_r|^{\tau} & \text{with probability } (1-\rho) \end{cases}$

- Where:
  - δ = random sign (trait increasing/decreasing)
  - τ = measures how the mean absolute effect of a mutation on the trait increases with the strength of selection
  - ρ = Probability that the trait effect is proportional to the selection coefficient: **Pleiotropy**!!
  - *s* = selection coefficient
  - s_r = random selection coefficient

Uricchio et al, Genome Research (2016)

Evolutionary Models Of Complex Disease



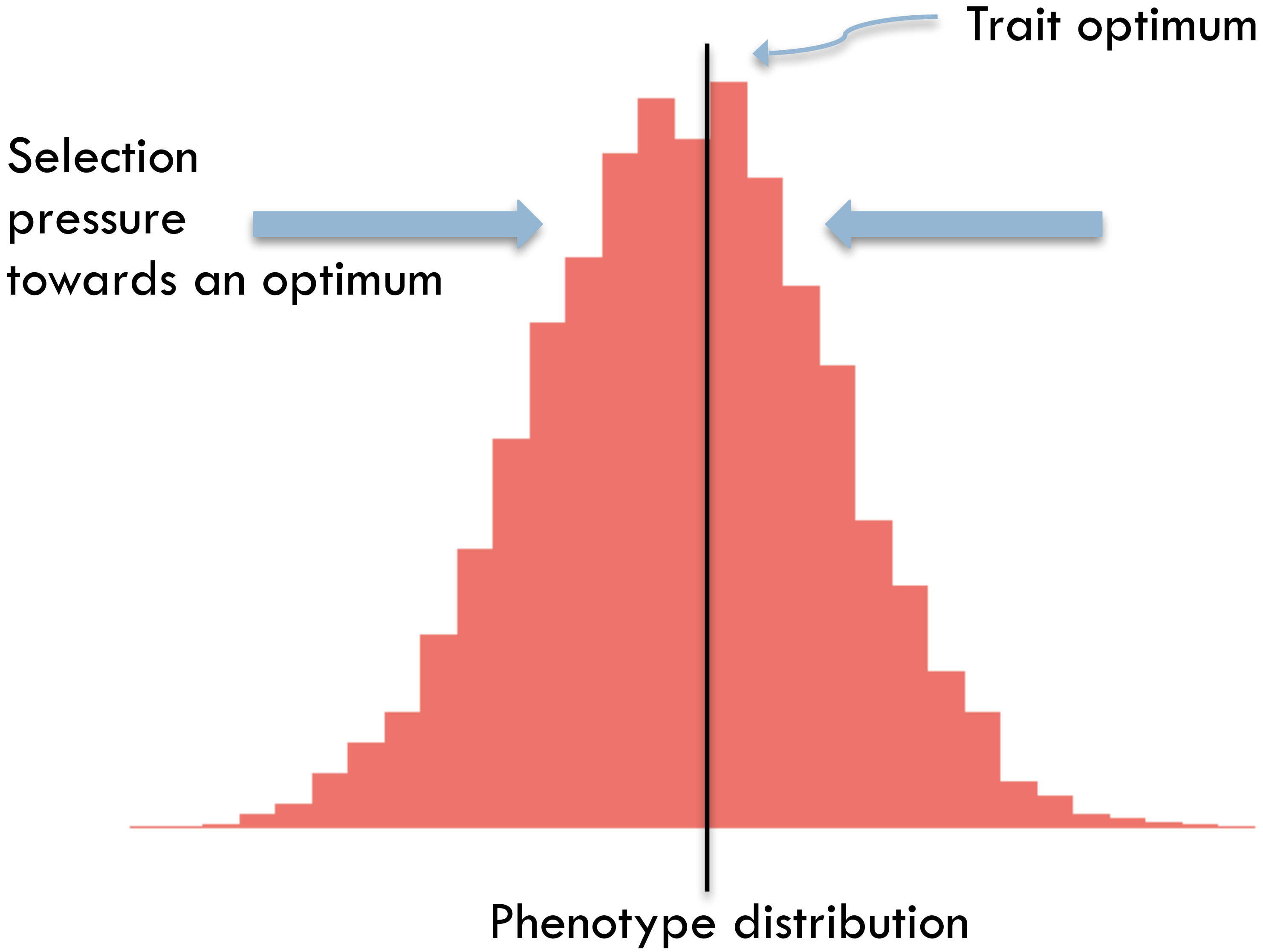
ρ: correlation(effect size, fitness) (Simons *et al*, 2014)

τ: transforms fitness effect to phenotype (Eyre-Walker, 2010)

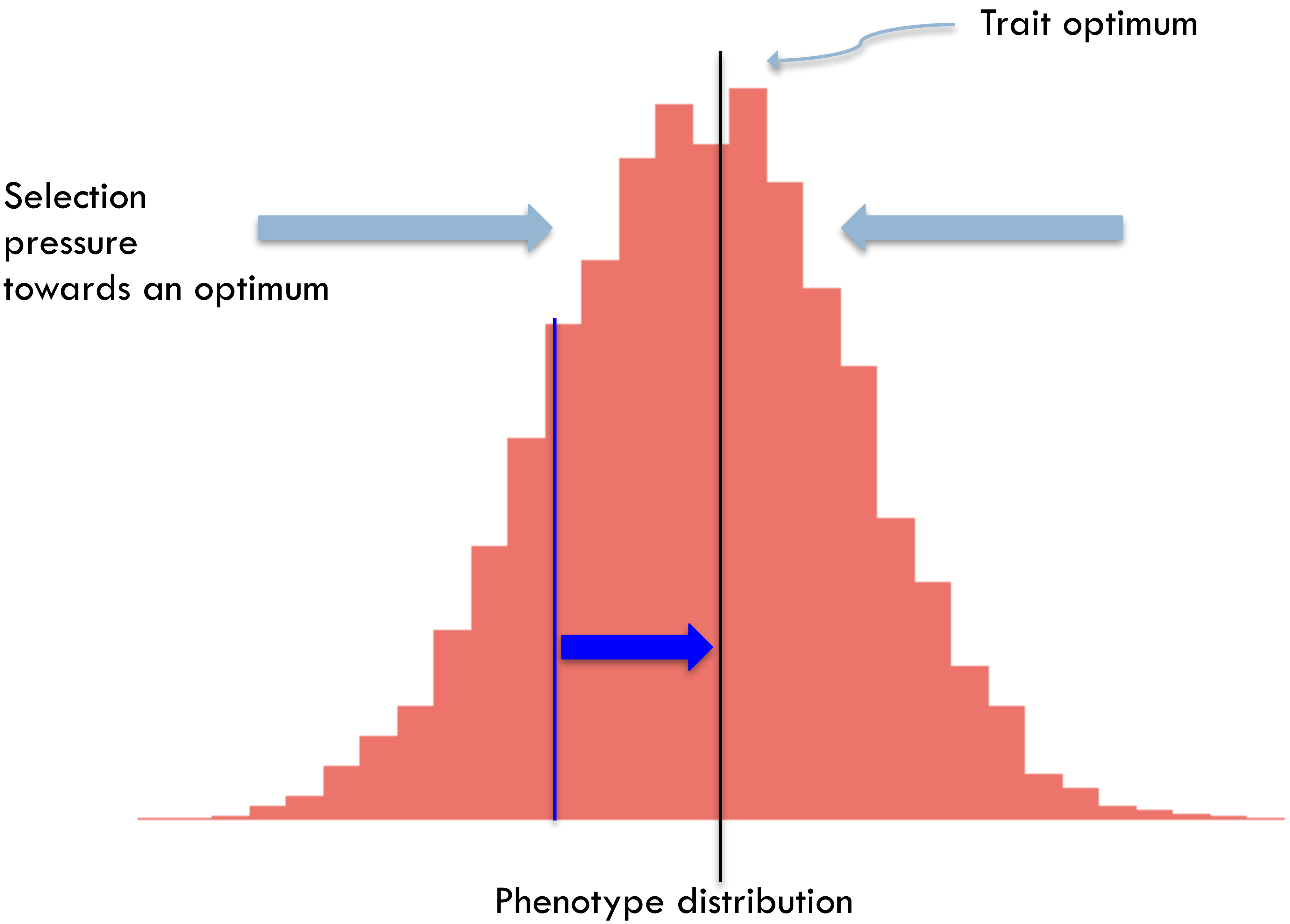
Pleiotropy: SNP impacts multiple phenotypes

Uricchio et al., Genome Research (2016)

Why should we think about evolution?



Stabilizing selection

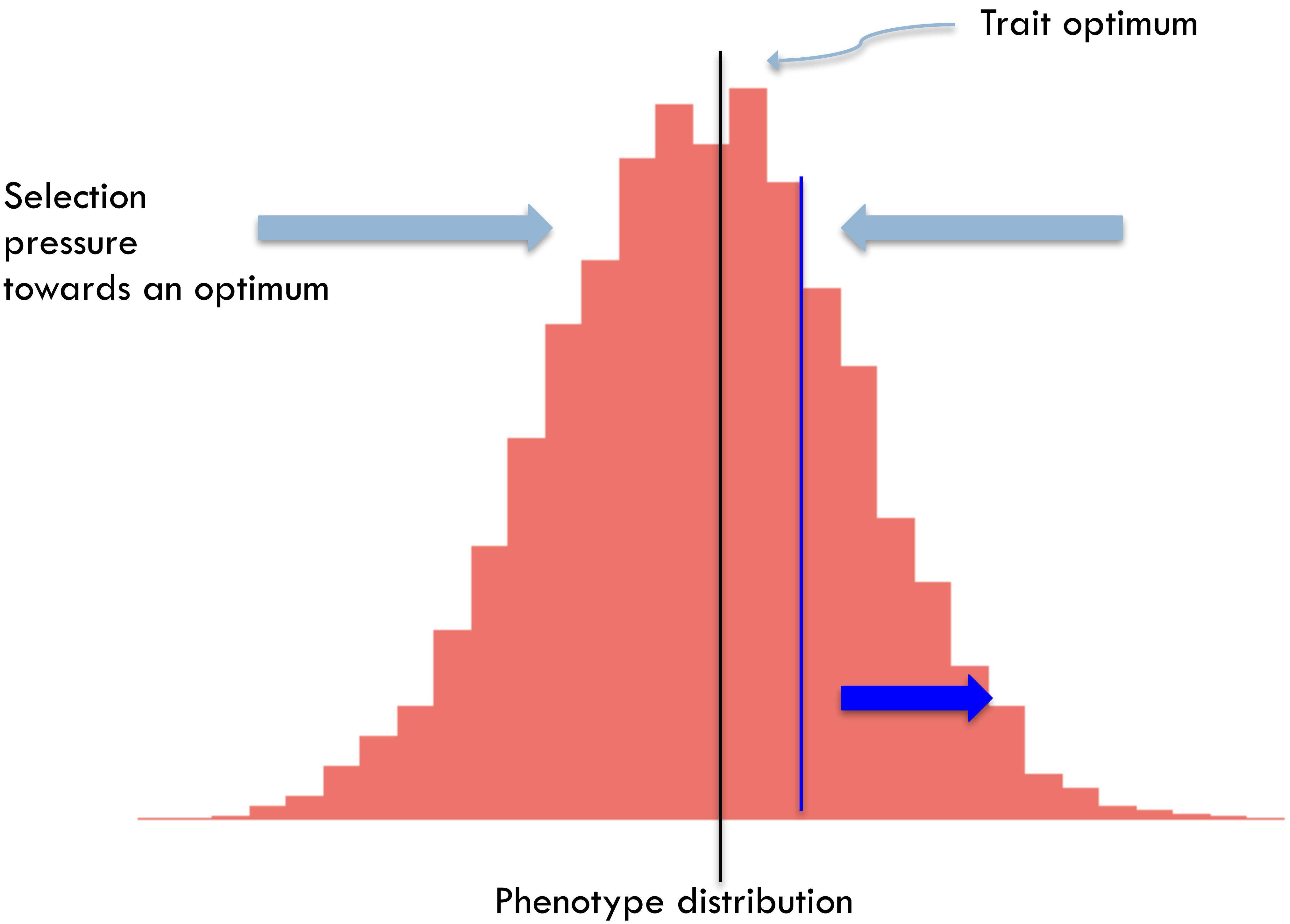


Trait optimum

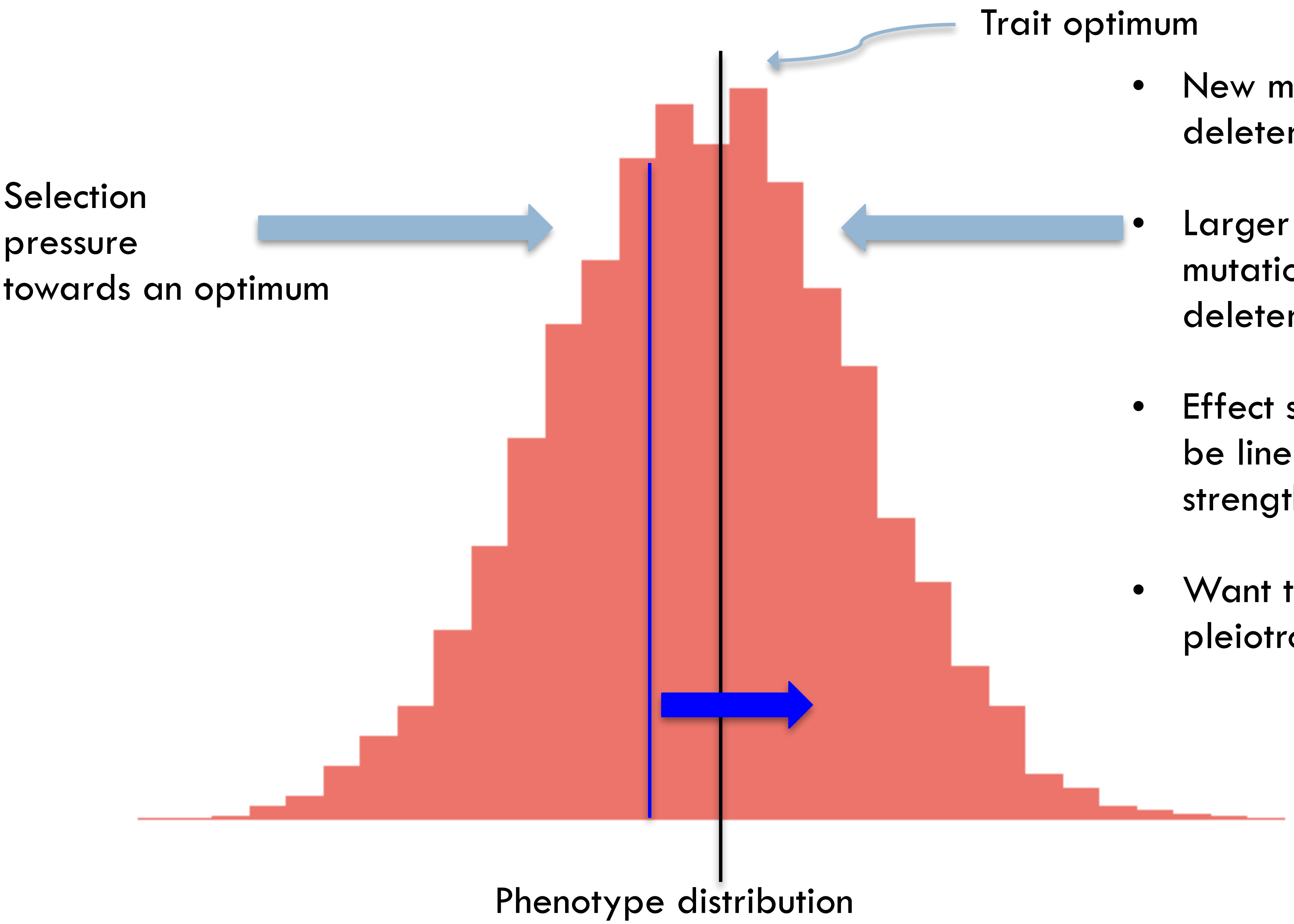
Selection
pressure
towards an optimum

Phenotype distribution

Stabilizing selection

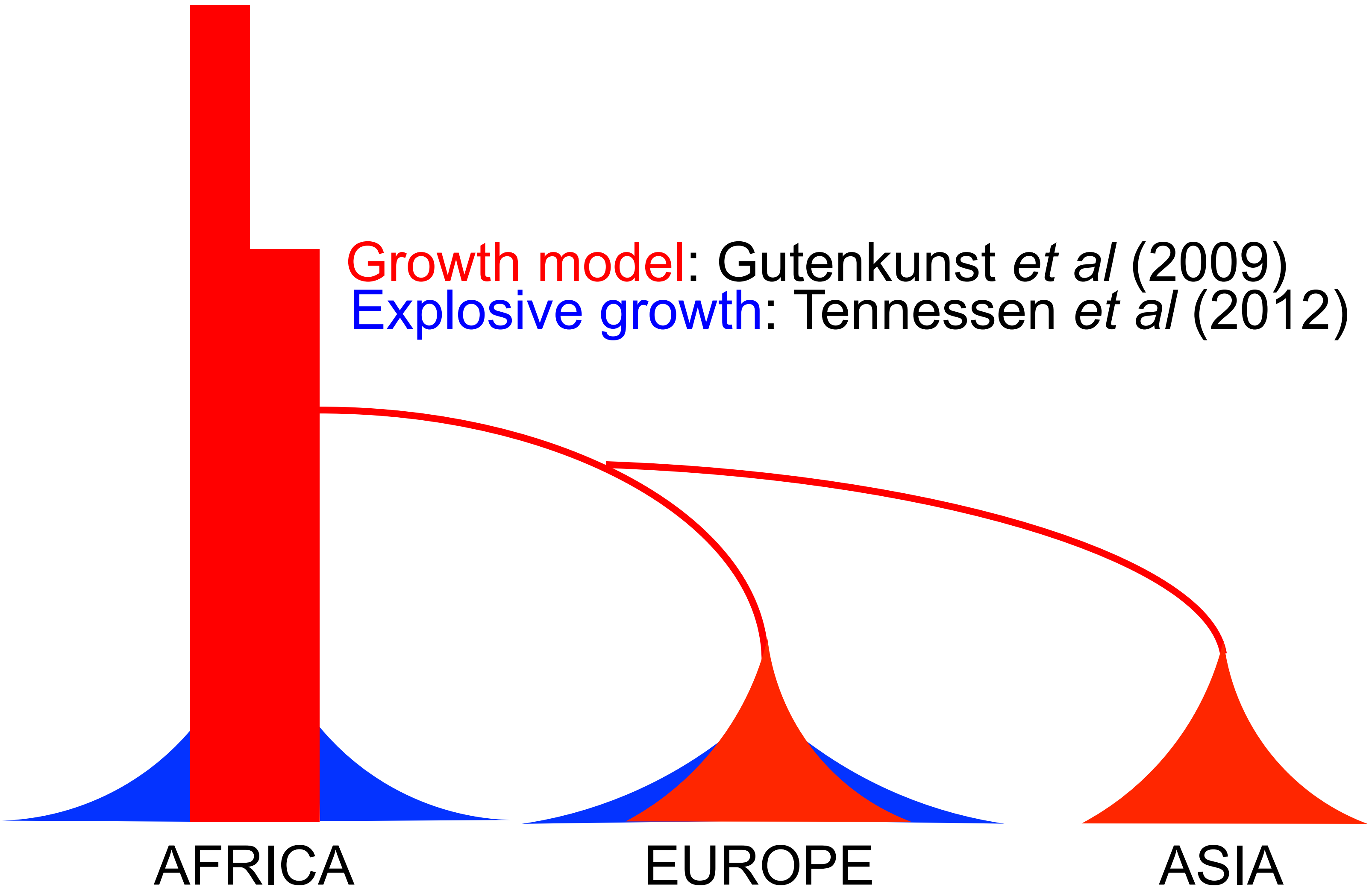


Stabilizing selection

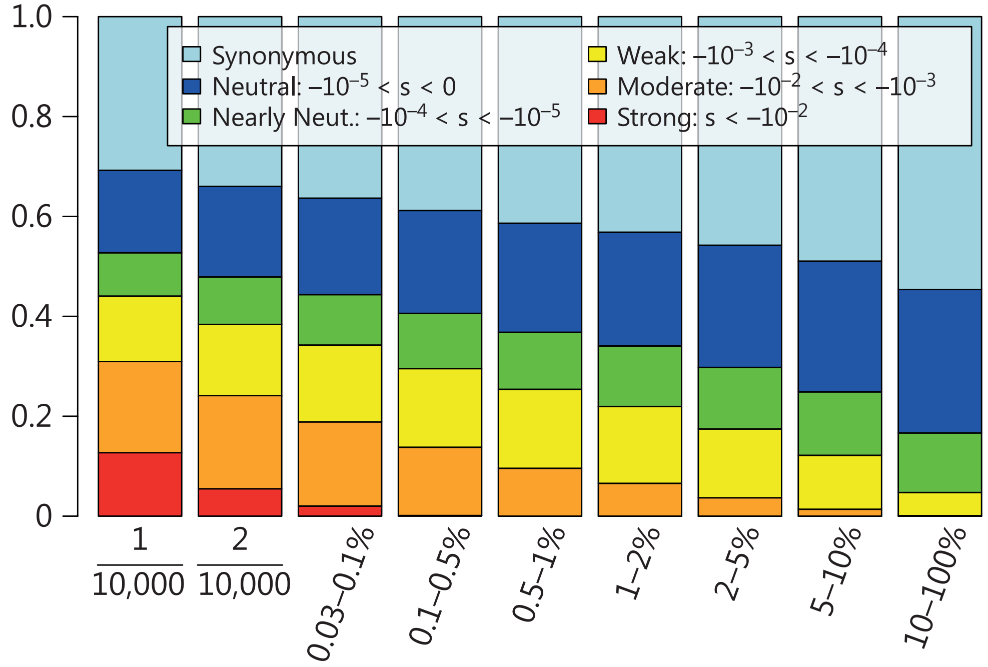


- New mutations deleterious
- Larger effect mutations are more deleterious
- Effect sizes may not be linear in selection strength
- Want to allow for pleiotropy

Human-specific demography and Selection


Growth model: Gutenkunst *et al* (2009)
Explosive growth: Tennessen *et al* (2012)
AFRICA
EUROPE
ASIA

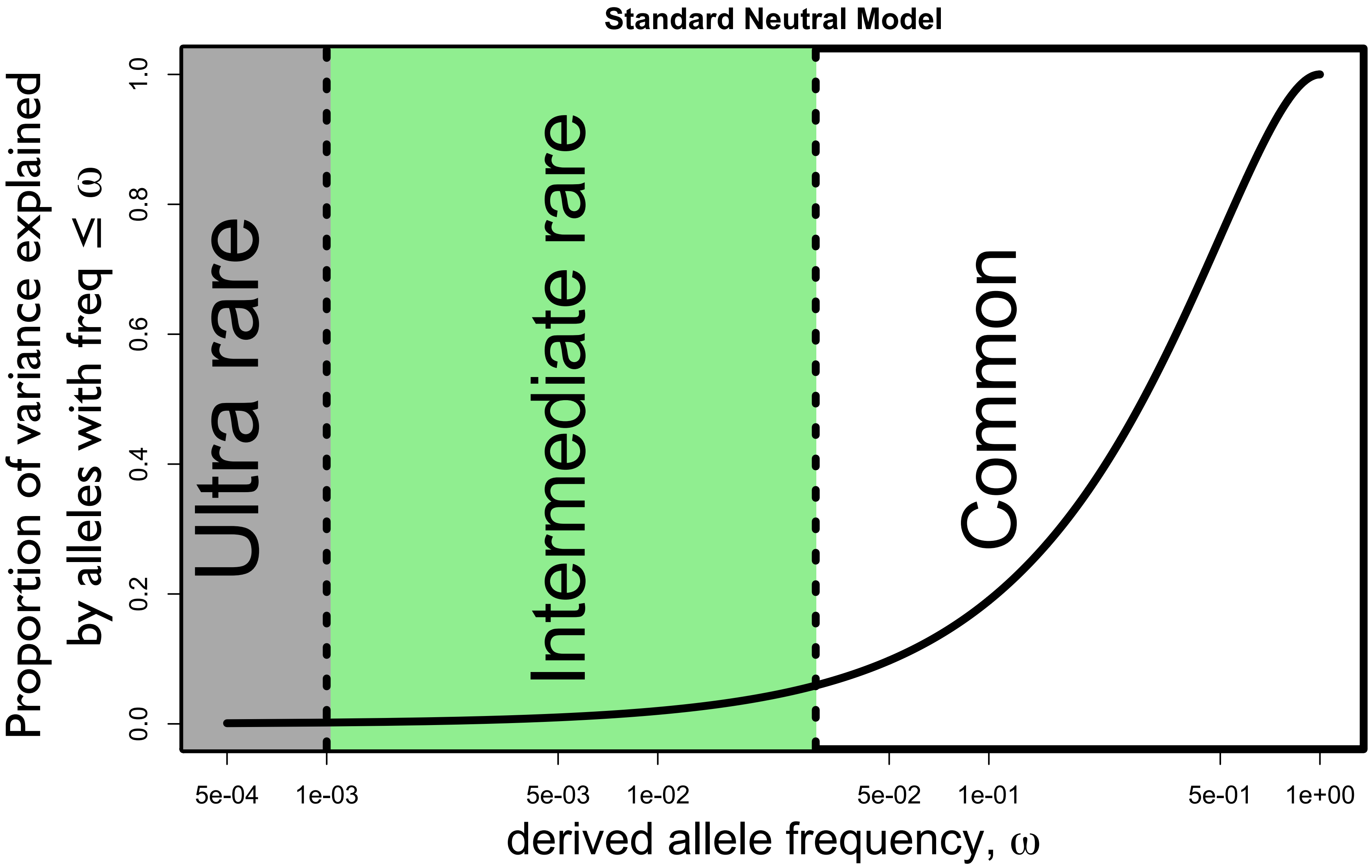
Fitness effects in non-coding DNA:
Torgerson *et al* (2009)



effect size = f(demography, natural selection)

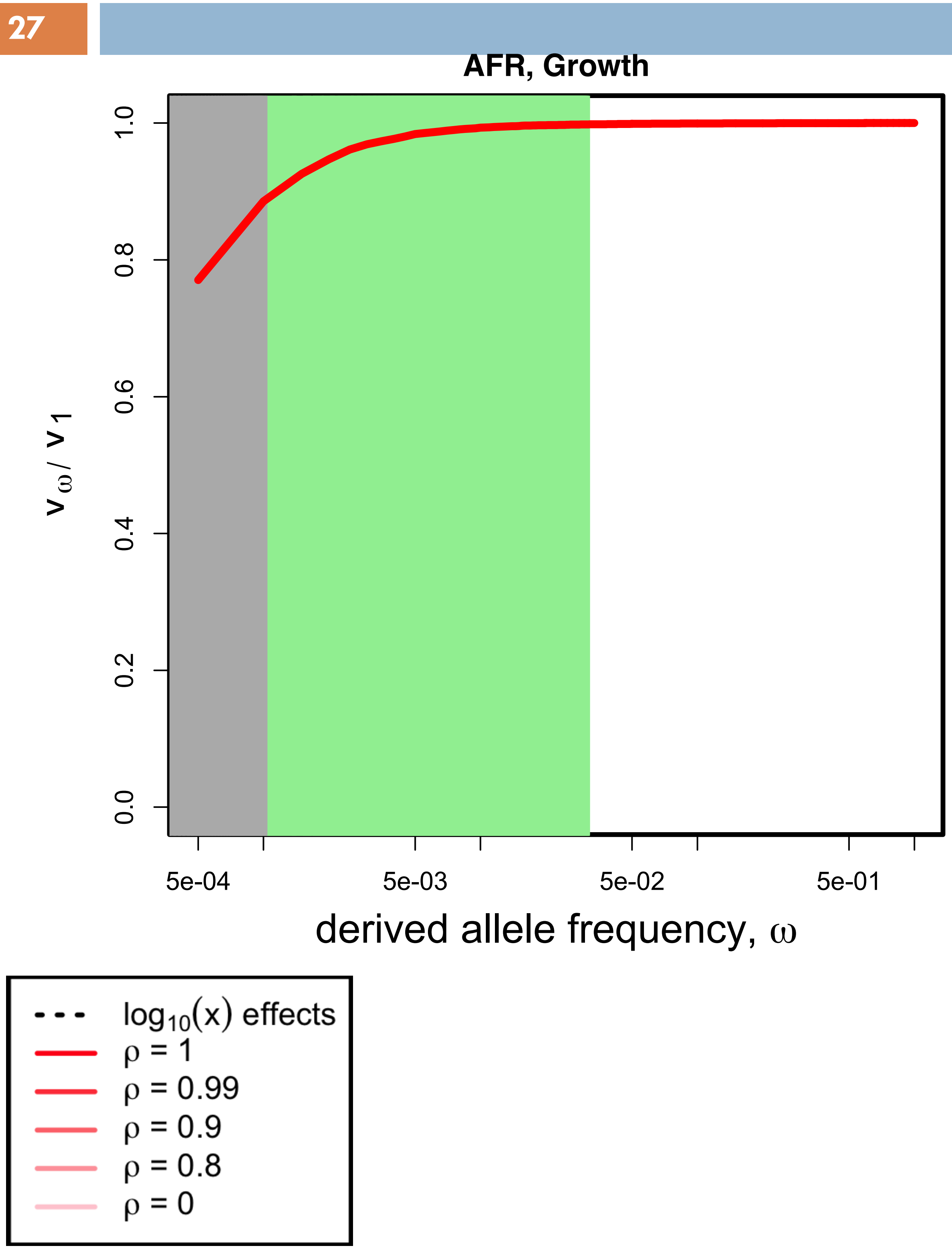
Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Neutral model: most variance explained by common alleles



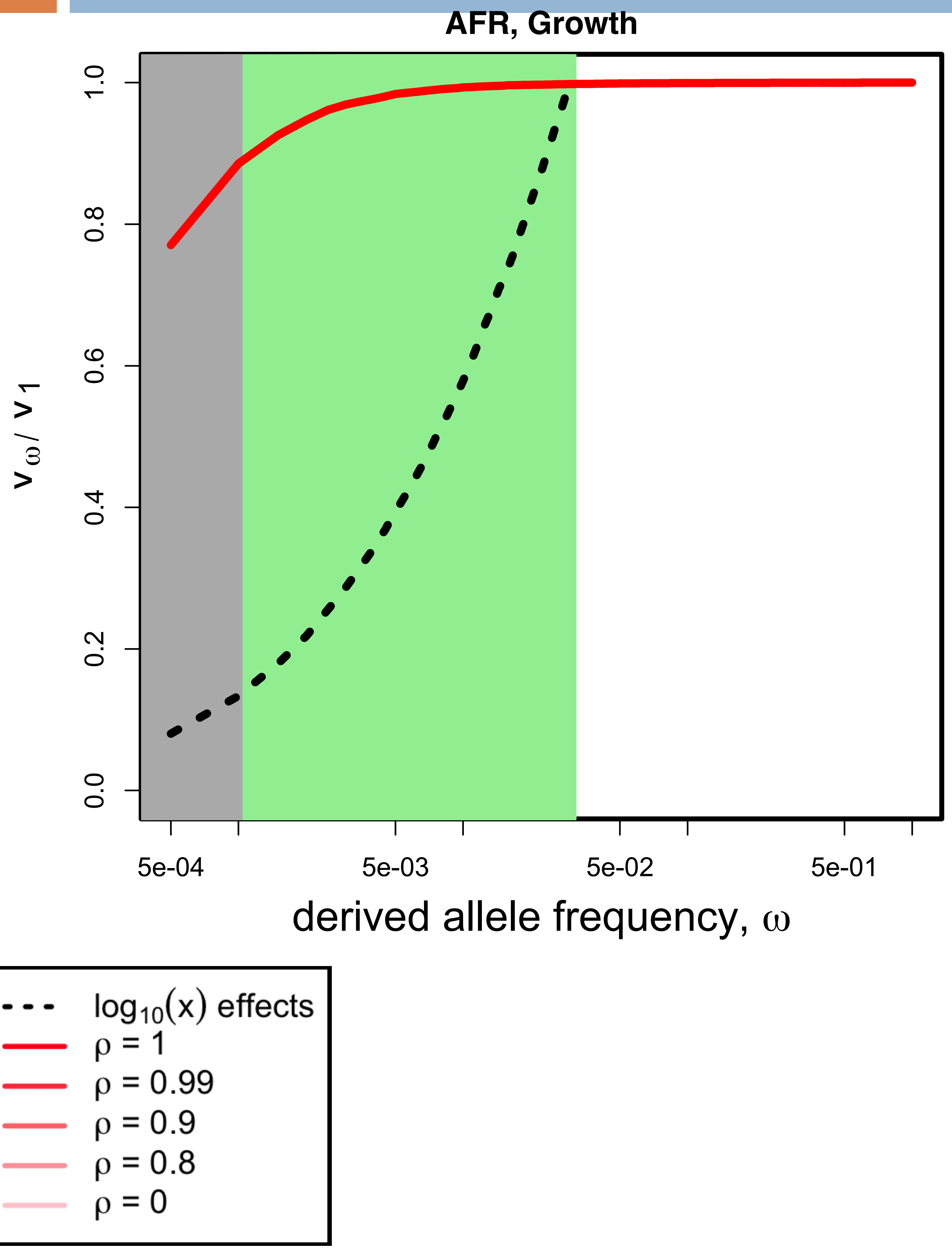
Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Genetic architecture is altered by selection and demography



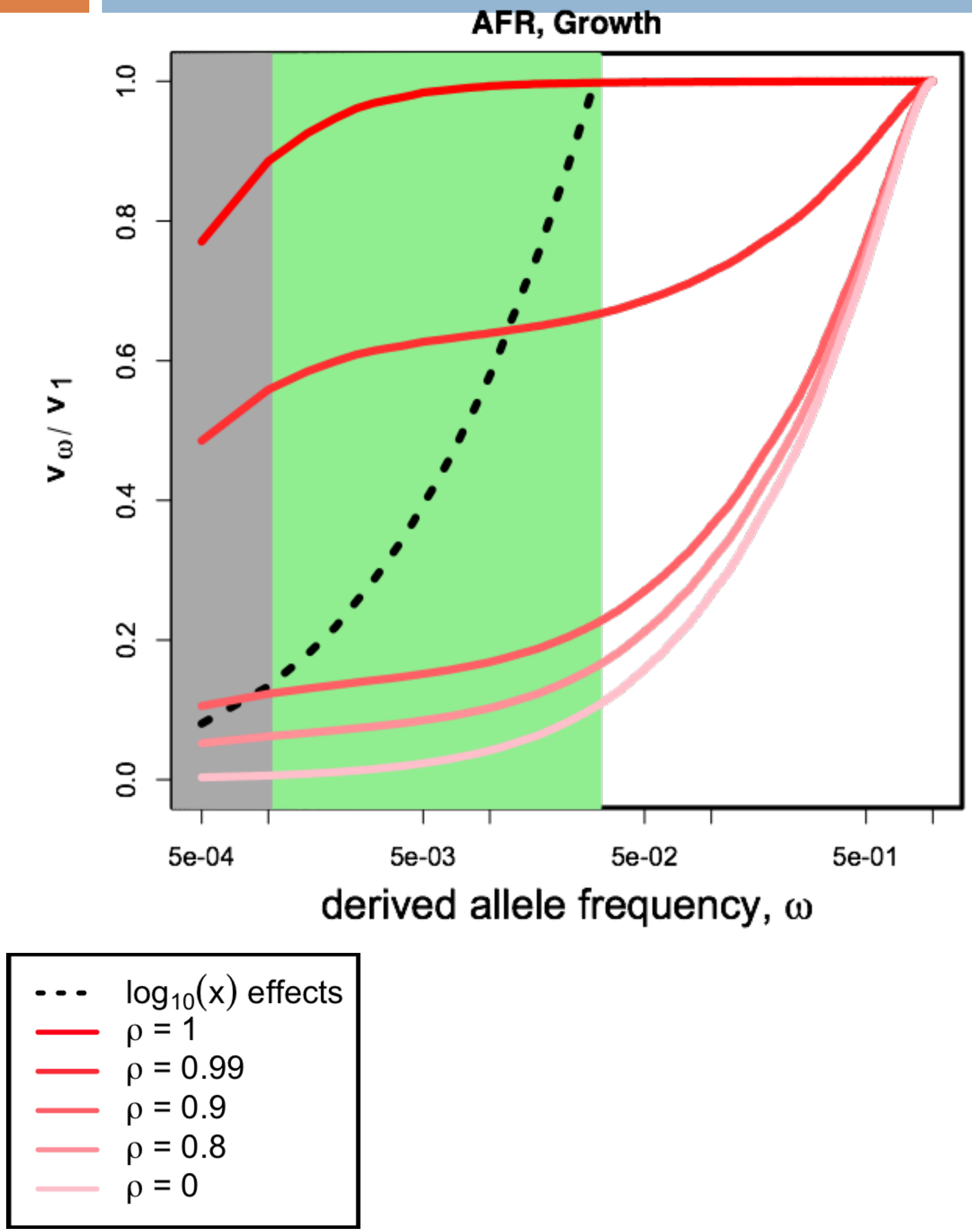
Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Genetic architecture is altered by selection and demography



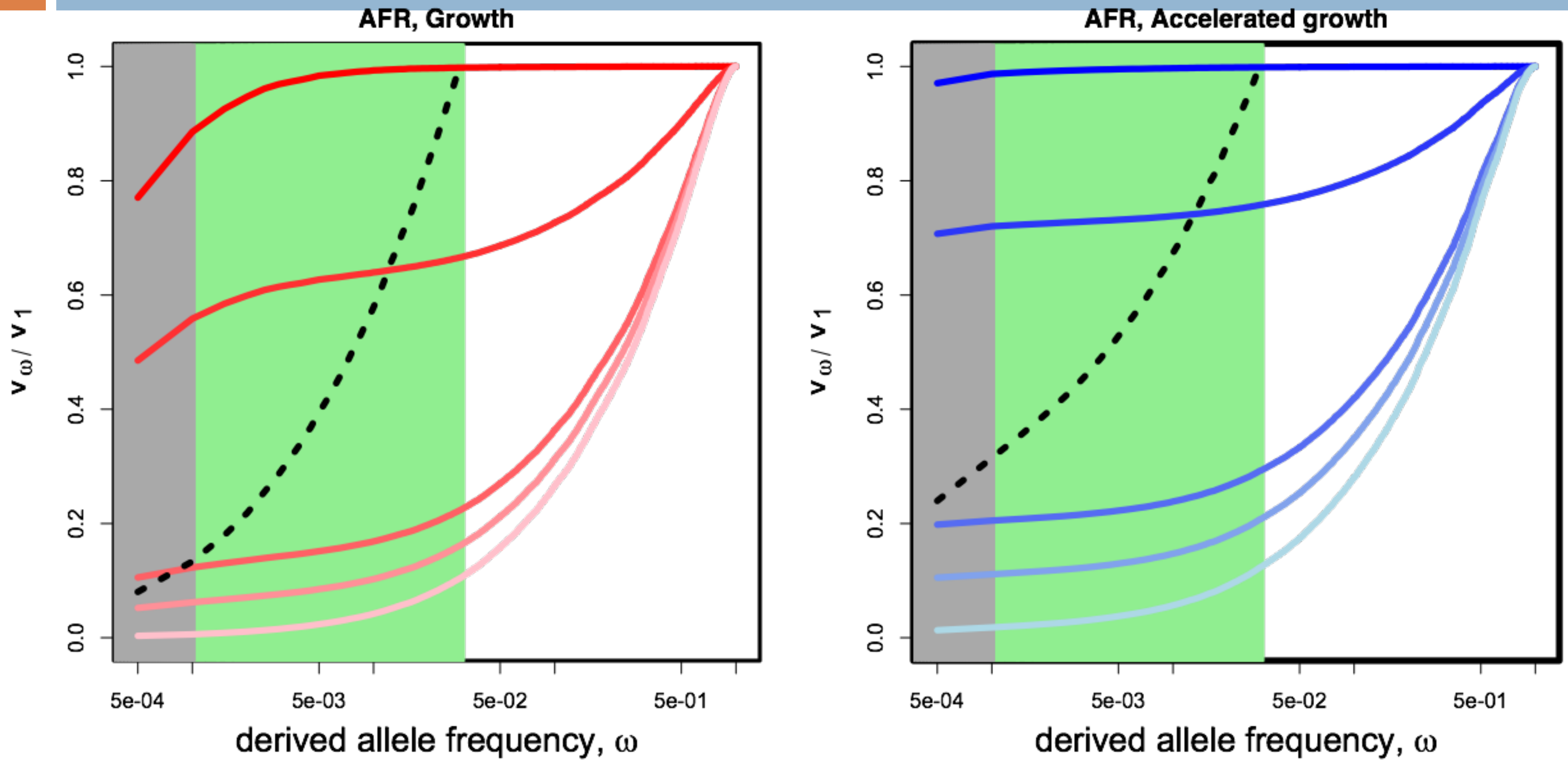
Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

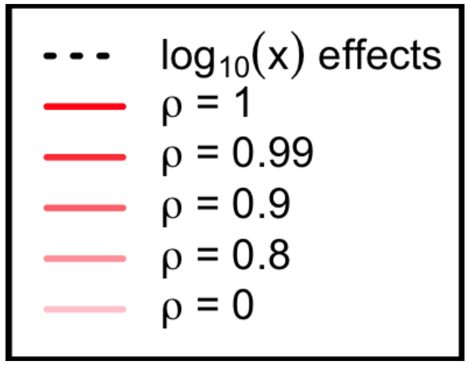
Genetic architecture is altered by selection and demography



Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Genetic architecture is altered by selection and demography



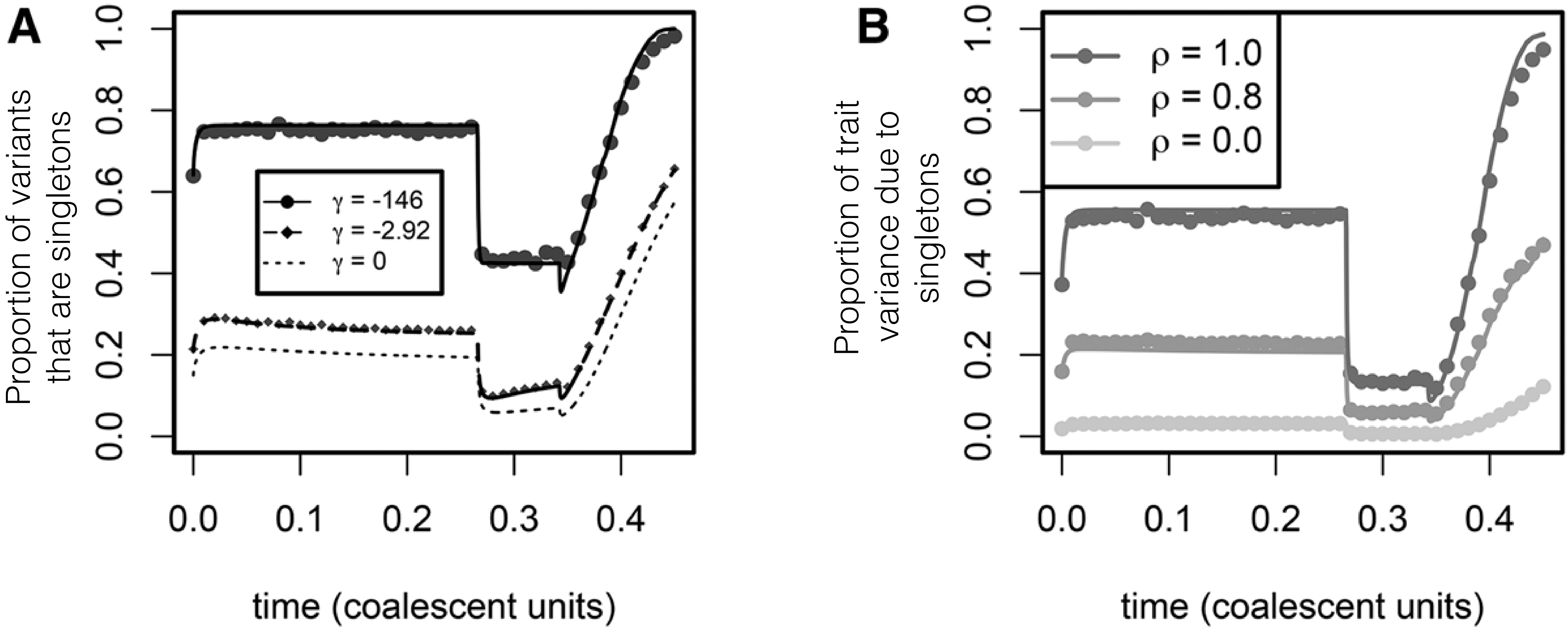


Implication: in some cases, largest effect alleles are very rare, so we may not detect them with GWAS!

Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Demography and selection matter!

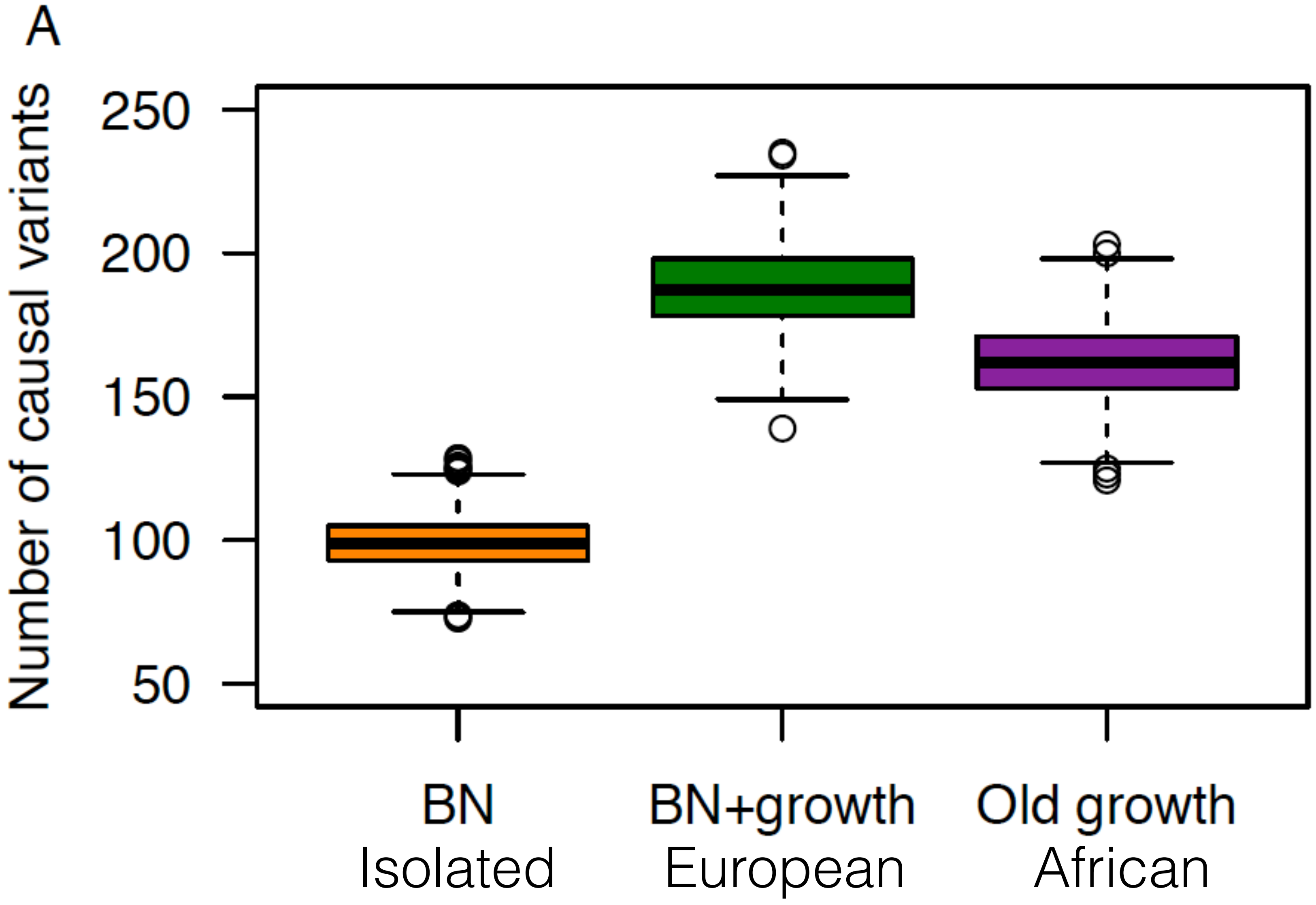
- As populations expand and contract, or strength of selection changes, the frequency spectrum responds.
- This can **and should** impact the genetic architecture of traits!



Uricchio, et al. *Genome Res* **26**, 863-873 (2016).

Demography and selection matter!

- Demography and selection also impacts the number of causal variants!



Lohmueller, *PLoS Genet* (2014).

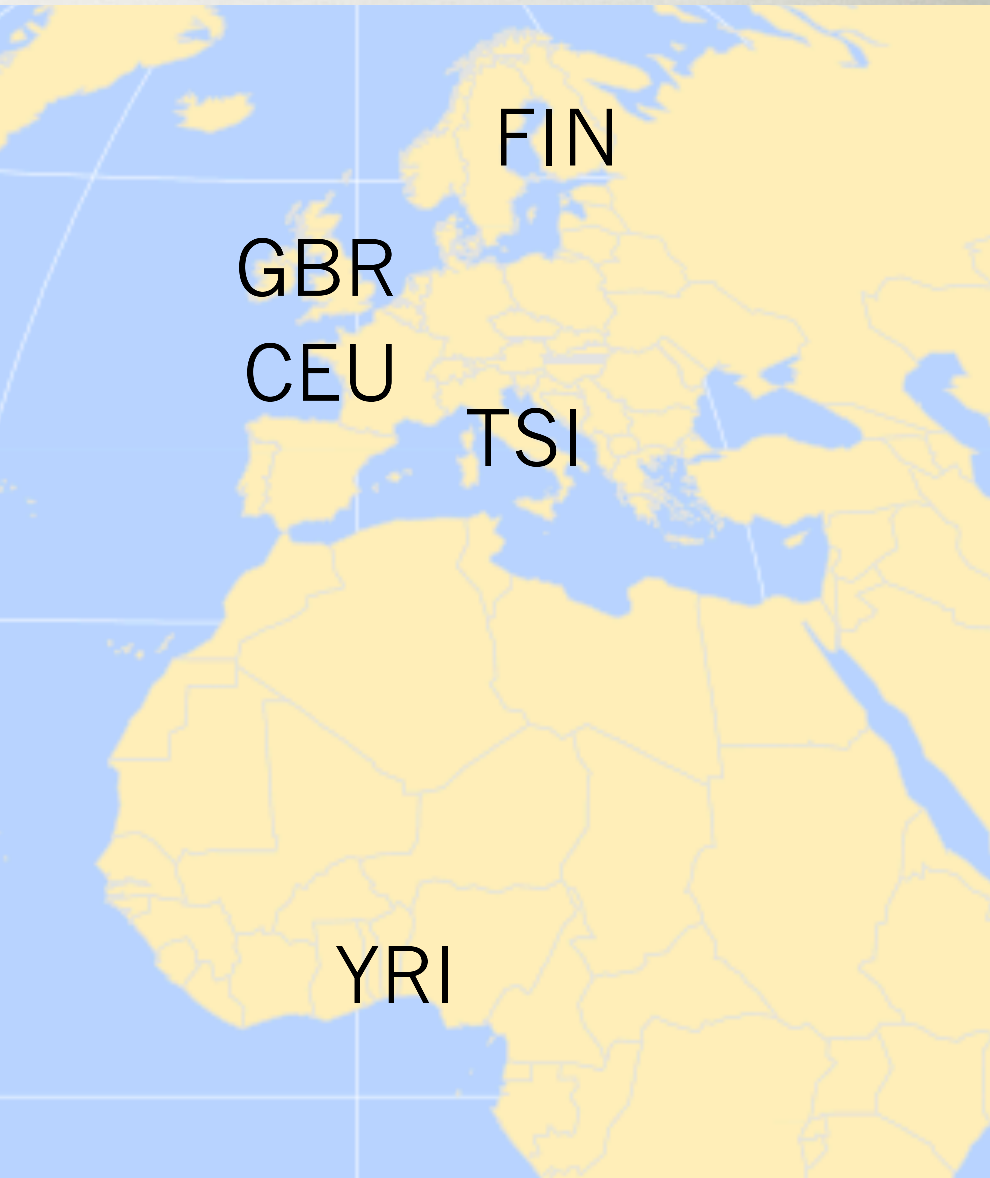
Open Questions

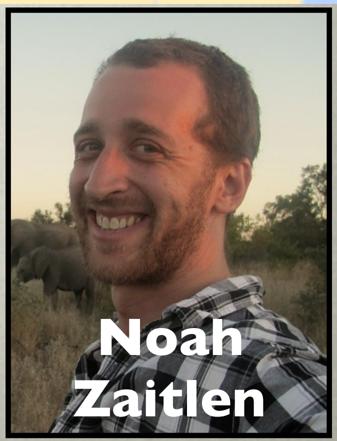
- What does does the genetic architecture of a complex trait really look like?
    - How many causal variants are there?
    - Proportion of effects from rare/common alleles?
    - Additive vs epistatic interactions?
    - Pleiotropy?

SEUVADIS CONSORTIUM

1000 Genomes
A Deep Catalog of Human Genetic Variation

- Large-scale RNA sequencing + WGS
  - 4 European populations
  - 360 individuals
  - low coverage WGS + high coverage exome: Phase 3.
  - RNA-seq: median depth 58.3M reads
    - Gene expression: log2 transformed, median centered, and quantile normalized.
    - 10,077 unique genes.




Noah Zaitlen

Hernandez, et al. (bioRxiv, 2019)

GEUVADIS CONSORTIUM

1000 Genomes
A Deep Catalog of Human Genetic Variation

- Large-scale RNA sequencing + WGS
  - 4 European populations
  - 360 individuals
  - low coverage WGS + high coverage exome: Phase 3.
  - RNA-seq: median depth 58.3M reads
    - Gene expression:
      log2 transformed, median centered, and quantile normalized.
    - 10,077 unique genes.

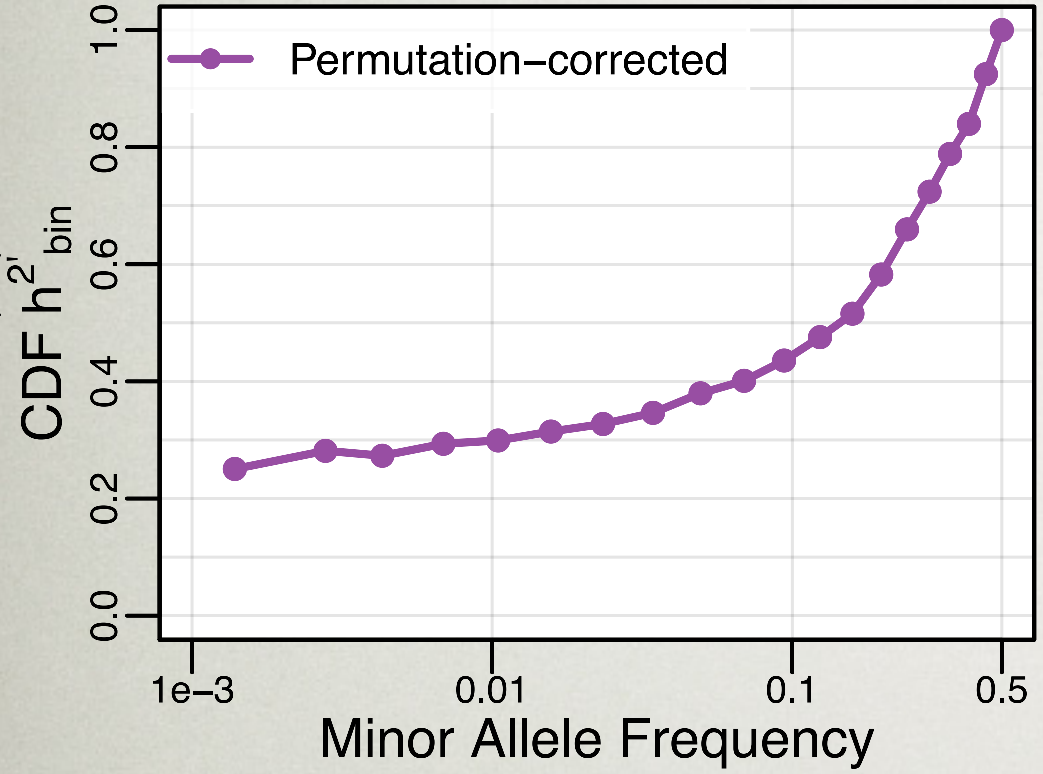
- Our sample size is **small**, but can we learn anything about the **genetic basis of complex traits from these 10k genes**?
- Let's analyze heritability of gene expression due to *cis* variation (within 1Mb of gene)

Estimating parameters

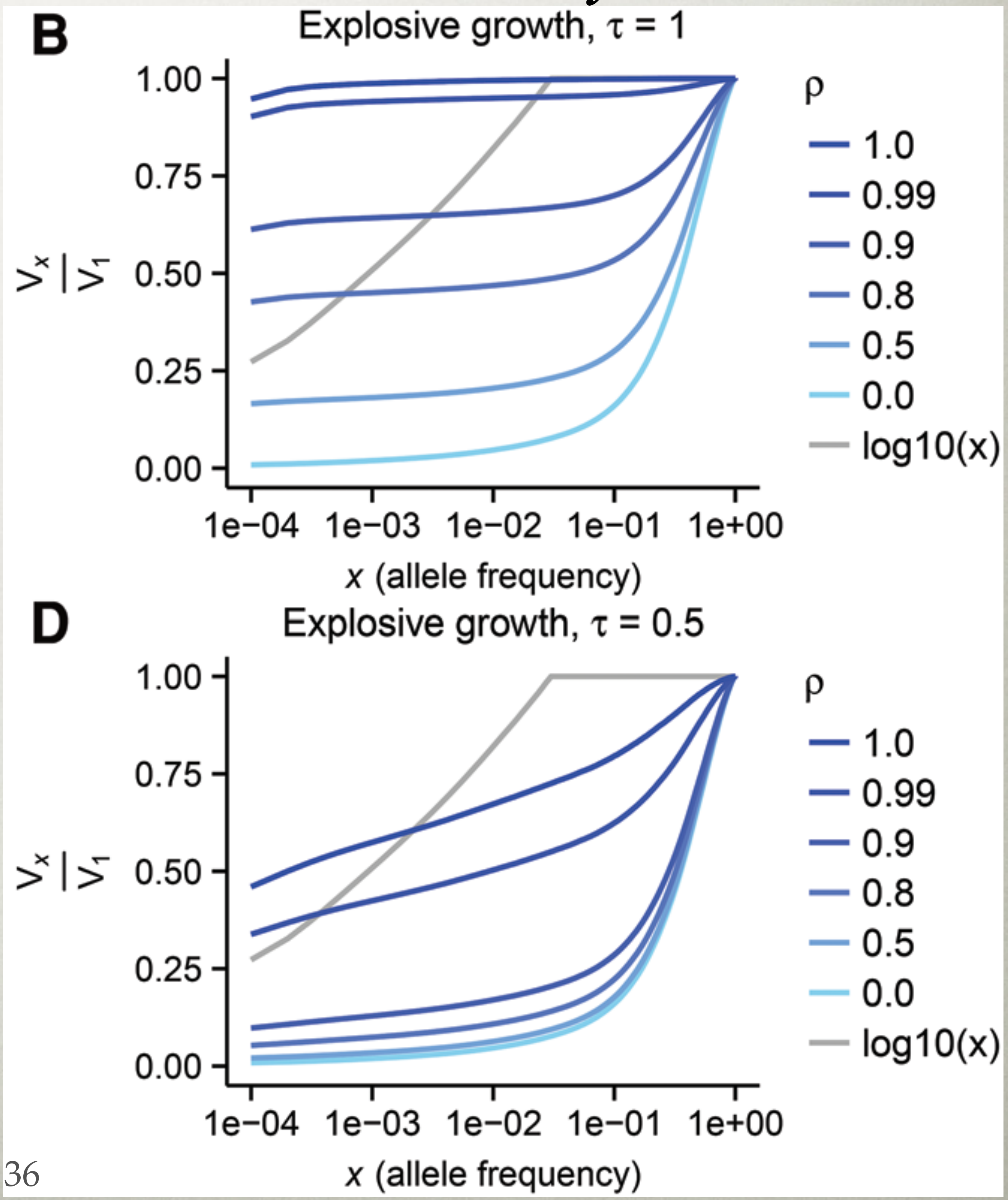
GEUVADIS CONSORTIUM

1000 Genomes
A Deep Catalog of Human Genetic Variation

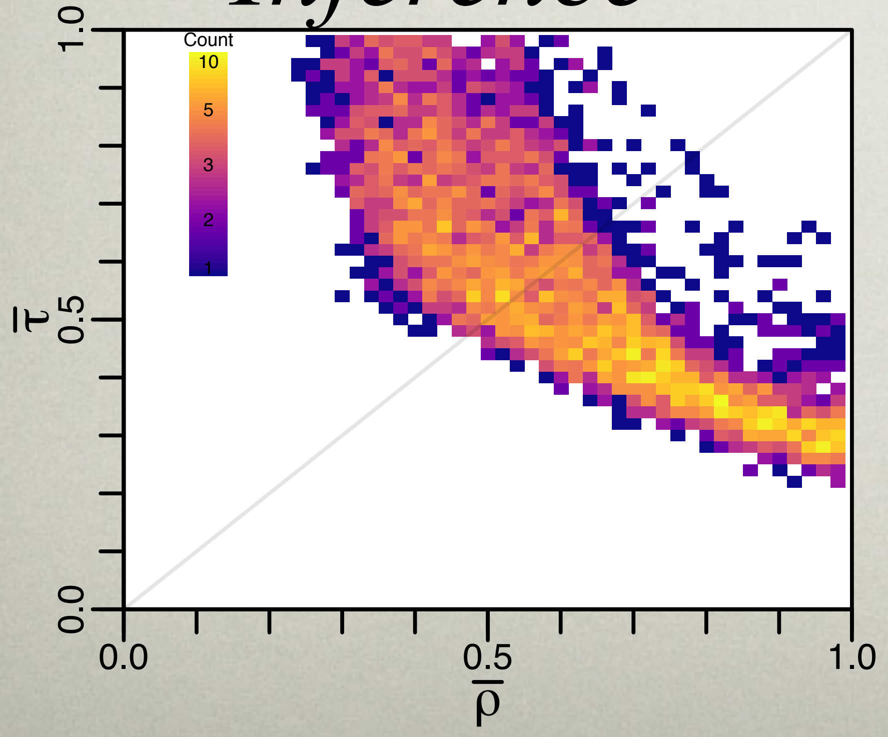
Data



Theory



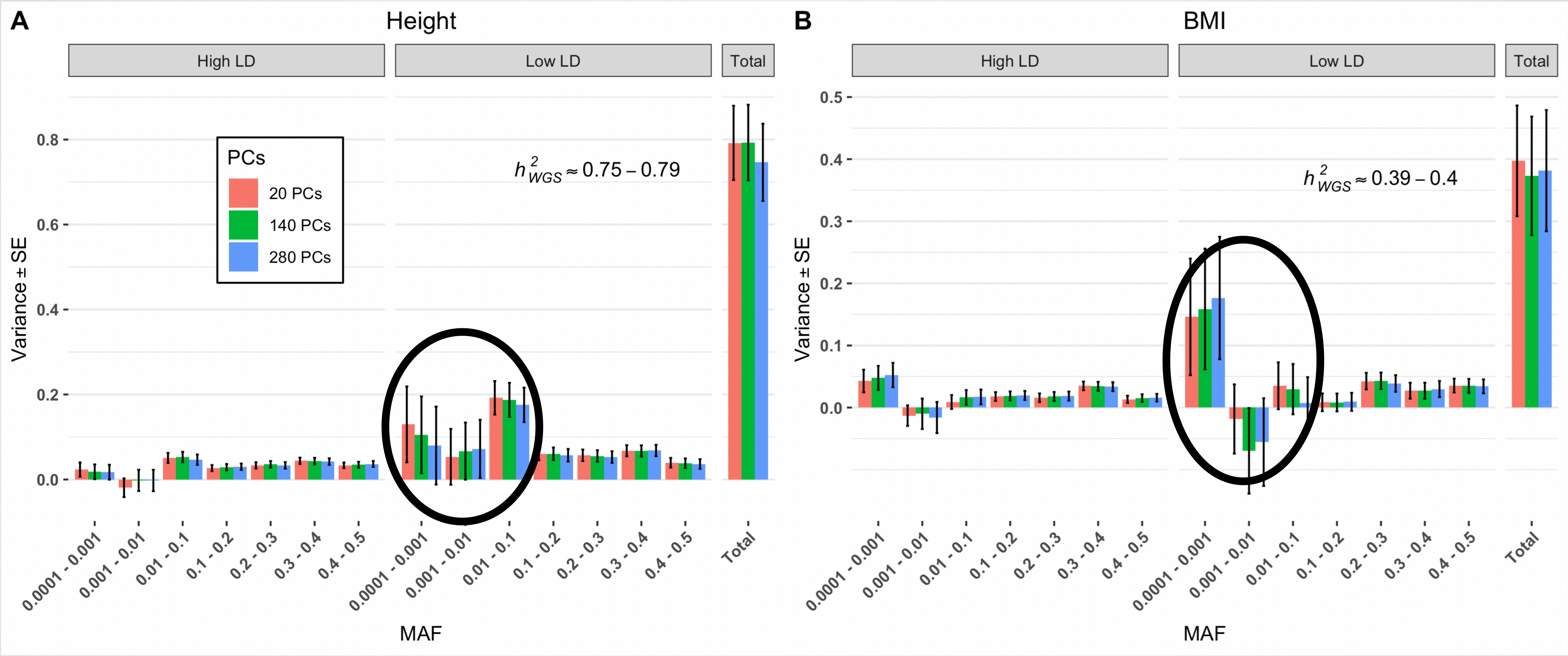
Inference



HUMAN HEIGHT AND BMI

n = 21,620 Individuals

Low MAF explains >50% of heritability



Wainschtein, et al. Recovery of trait heritability from whole genome sequence data. **bioRxiv**.

Simulating Genetic Architectures and Inferring Heritability

Goals

- To learn what heritability is
- To learn how to calculate it from unrelated samples
- To learn how to simulate phenotypes & evaluate performance of the test

What is heritability?

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

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

$$\sigma_{G}^{2} = \sigma_{A}^{2} + \sigma_{D}^{2} + \sigma_{I}^{2}$$

$$\text{Narrow Sense}: \ h^{2} = \frac{\sigma_{A}^{2}}{\sigma_{\varphi}^{2}} \qquad \text{Broad Sense}: \ H^{2} = \frac{\sigma_{G}^{2}}{\sigma_{\varphi}^{2}}$$

How do we estimate h^2?

- We will focus on Haseman-Elston (HE) regression
- Very simply:
  - Let p be the covariance in phenotypes across all individuals
  - Let g be the covariance in genotypes across all individuals
  - h^2 = the correlation between p and g!!

How do we simulate phenotypes?

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

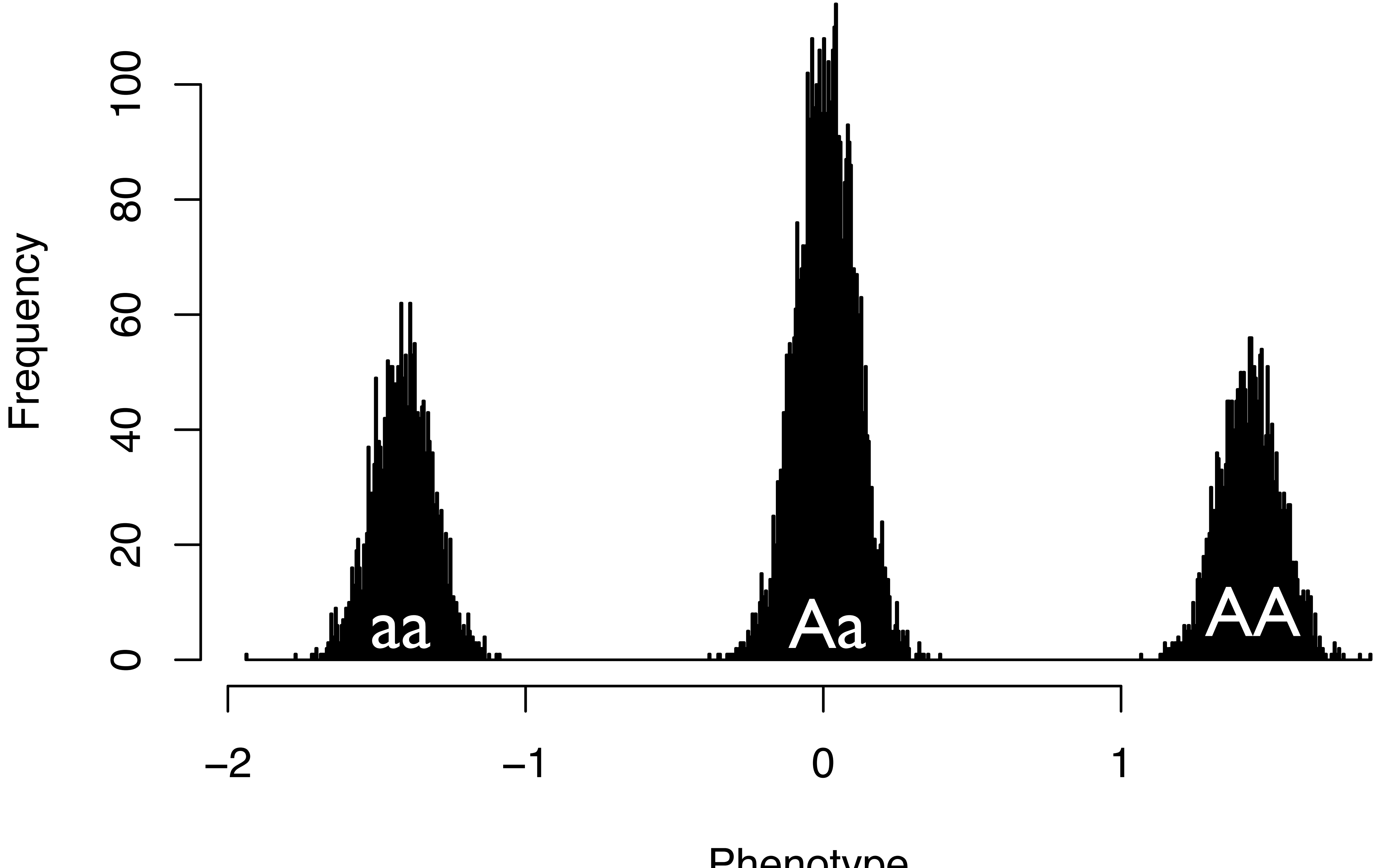
$$\sigma_{G}^{2} = \sigma_{A}^{2} + \sigma_{D}^{2} + \sigma_{I}^{2}$$

- The basic model of phenotypes we will assume is an additive model

- We will assume that environmental noise is $\sigma_{\varepsilon}^{2} \sim N(0,\sigma^{2})$

How do we simulate phenotypes?

- The genetic effect depends on causal variation!
- With 1 causal locus and little environmental noise:



How do we simulate phenotypes?

- The genetic effect depends on causal variation!
- How much environmental noise is there?
- It depends on your desired level of h^2!

How do we simulate phenotypes?

- We are going to do the simulations in **`R`**!!
- Open terminal/command prompt and type:
  - **`Rscript  HEplay.R`**
- If Rscript does not work on your computer, you can open R, and move to HEplay directory and type:
  - **`source(“HEplay.R”)`**

How do we simulate phenotypes?

- It will produce output like this:

```
rhernandez$ Rscript HEplay.R
Read 7570 items
Read 2725200 items
0.4468355 (mean = 0.4468355)
0.59926 (mean = 0.5230477)
0.6873345 (mean = 0.57781)
0.3375272 (mean = 0.5177393)
0.4301956 (mean = 0.5002305)
0.5716429 (mean = 0.5121326)
0.8160635 (mean = 0.5555513)
0.6663577 (mean = 0.5694021)
0.3248046 (mean = 0.5422246)
0.584494 (mean = 0.5464515)
0.4031187 (mean = 0.5334213)
0.6347714 (mean = 0.5418671)
0.3799034 (mean = 0.5294084)
0.3614569 (mean = 0.5174118)
0.4317423 (mean = 0.5117005)
0.6364826 (mean = 0.5194994)
0.5425433 (mean = 0.5208549)
0.5204382 (mean = 0.5208318)
0.6647941 (mean = 0.5284088)
0.6268889 (mean = 0.5333328)
0.5361605 (mean = 0.5334674)
0.5609872 (mean = 0.5347183)
0.650662 (mean = 0.5397593)
0.5030965 (mean = 0.5382317)
0.4885729 (mean = 0.5362454)
True h2 = 0.5
mean(estimated h2) +- 2SE = 0.5362454 +- 0.04996943
Relative Bias = 0.07249074
```

How do we simulate phenotypes?

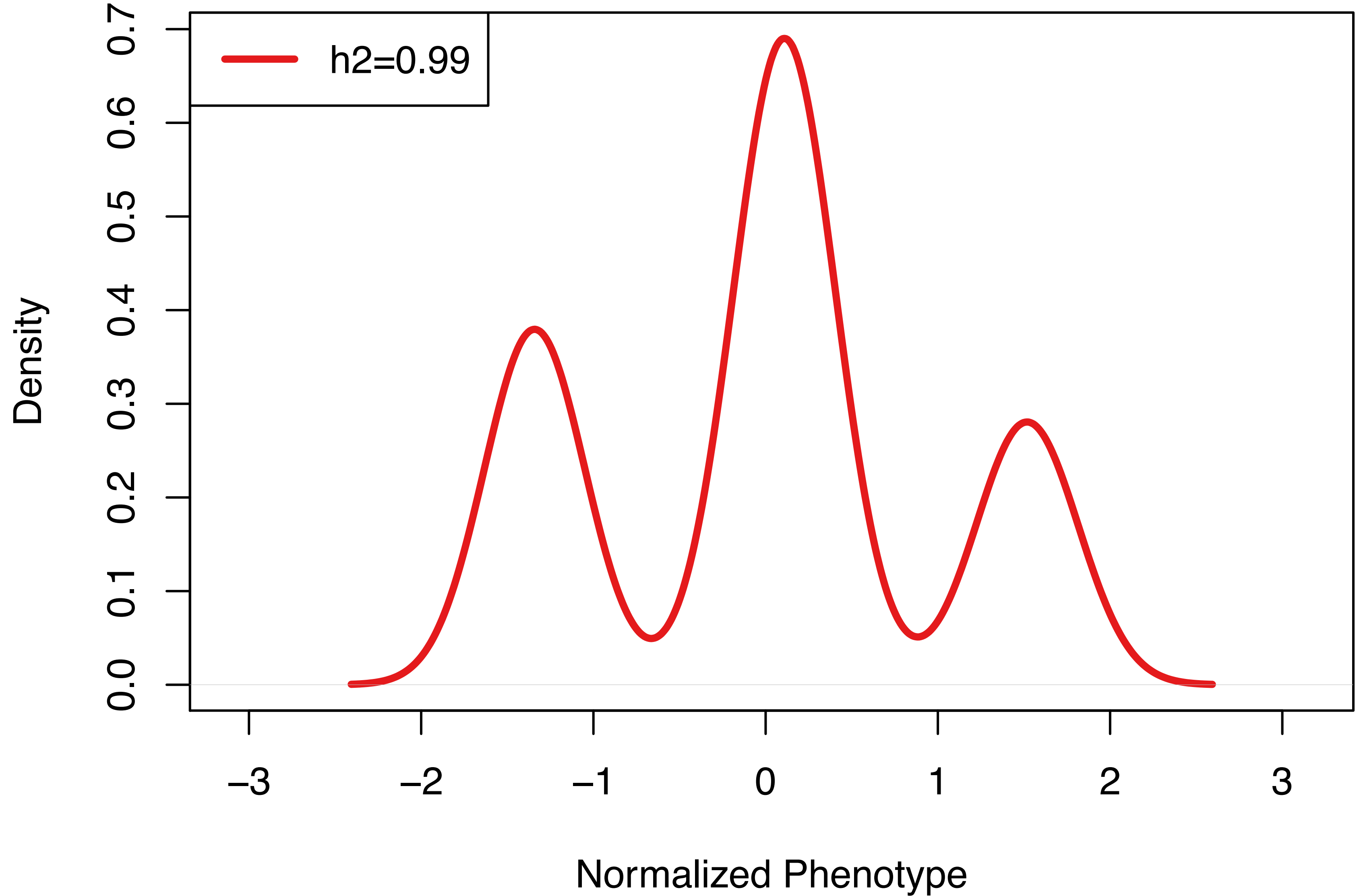
- It will produce output like this:

```
rhernandez$ Rscript HEplay.R
Read 7570 items
Read 2725200 items
0.4468355 (mean = 0.4468355)
0.59926 (mean = 0.5230477)
0.6873345 (mean = 0.57781)
0.3375272 (mean = 0.5177393)
0.4301956 (mean = 0.5002305)
0.5716429 (mean = 0.5121326)
0.8160635 (mean = 0.5555513)
0.6663577 (mean = 0.5694021)
0.3248046 (mean = 0.5422246)
0.584494 (mean = 0.5464515)
0.4031187 (mean = 0.5334213)
0.6347714 (mean = 0.5418671)
0.3799034 (mean = 0.5294084)
0.3614569 (mean = 0.5174118)
0.4317423 (mean = 0.5117005)
0.6364826 (mean = 0.5194994)
0.5425433 (mean = 0.5208549)
0.5204382 (mean = 0.5208318)
0.6647941 (mean = 0.5284088)
0.6268889 (mean = 0.5333328)
0.5361605 (mean = 0.5334674)
0.5609872 (mean = 0.5347183)
0.650662 (mean = 0.5397593)
0.5030965 (mean = 0.5382317)
0.4885729 (mean = 0.5362454)
True h2 = 0.5
mean(estimated h2) +- 2SE = 0.5362454 +- 0.04996943
Relative Bias = 0.07249074
```

Who got the largest/ smallest value?

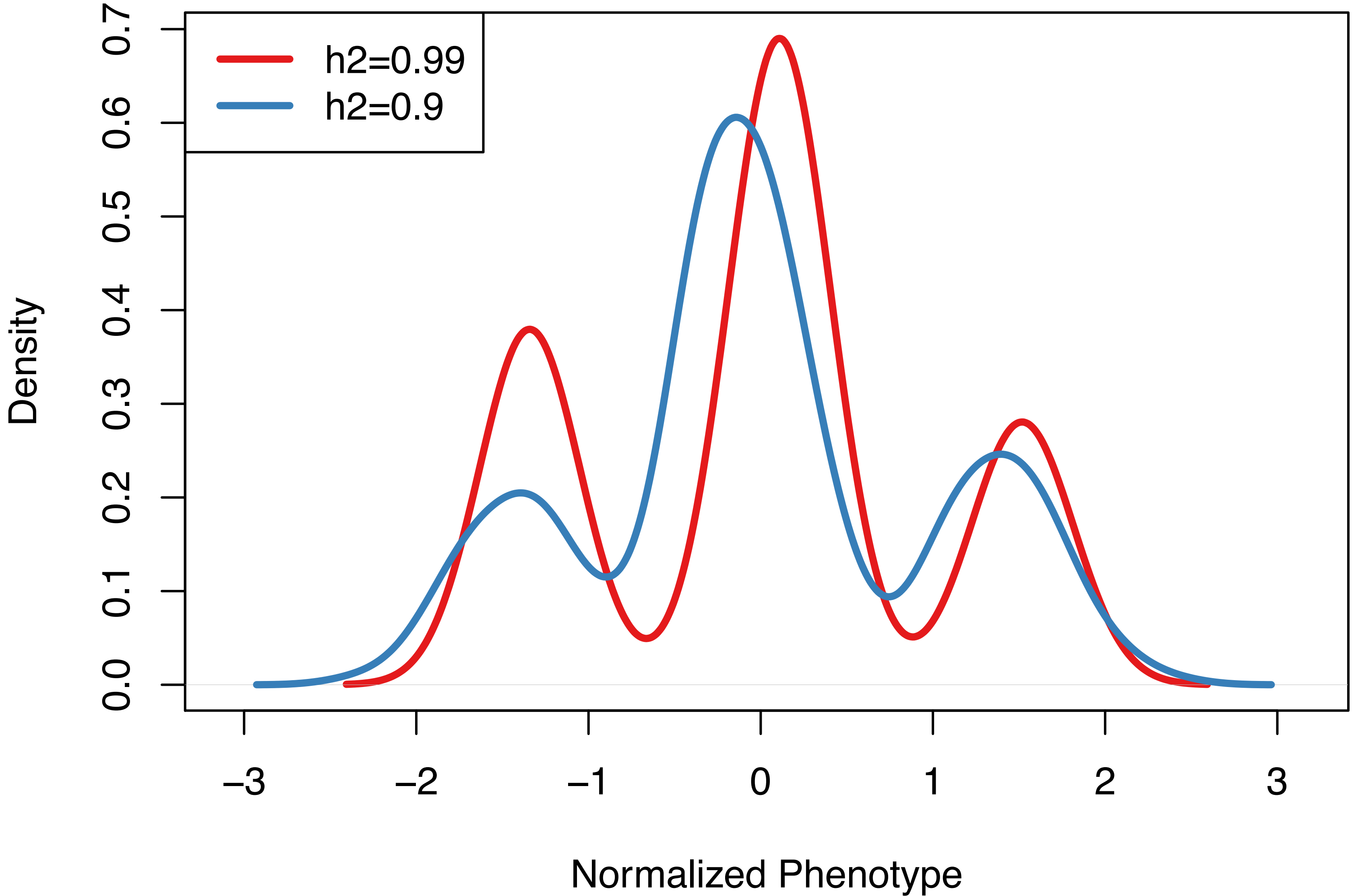
How do we simulate phenotypes?

- Simplest model:
  - There is 1 causal SNP.
  - Reference allele has no effect, but alternate allele has “some non-zero effect size”.



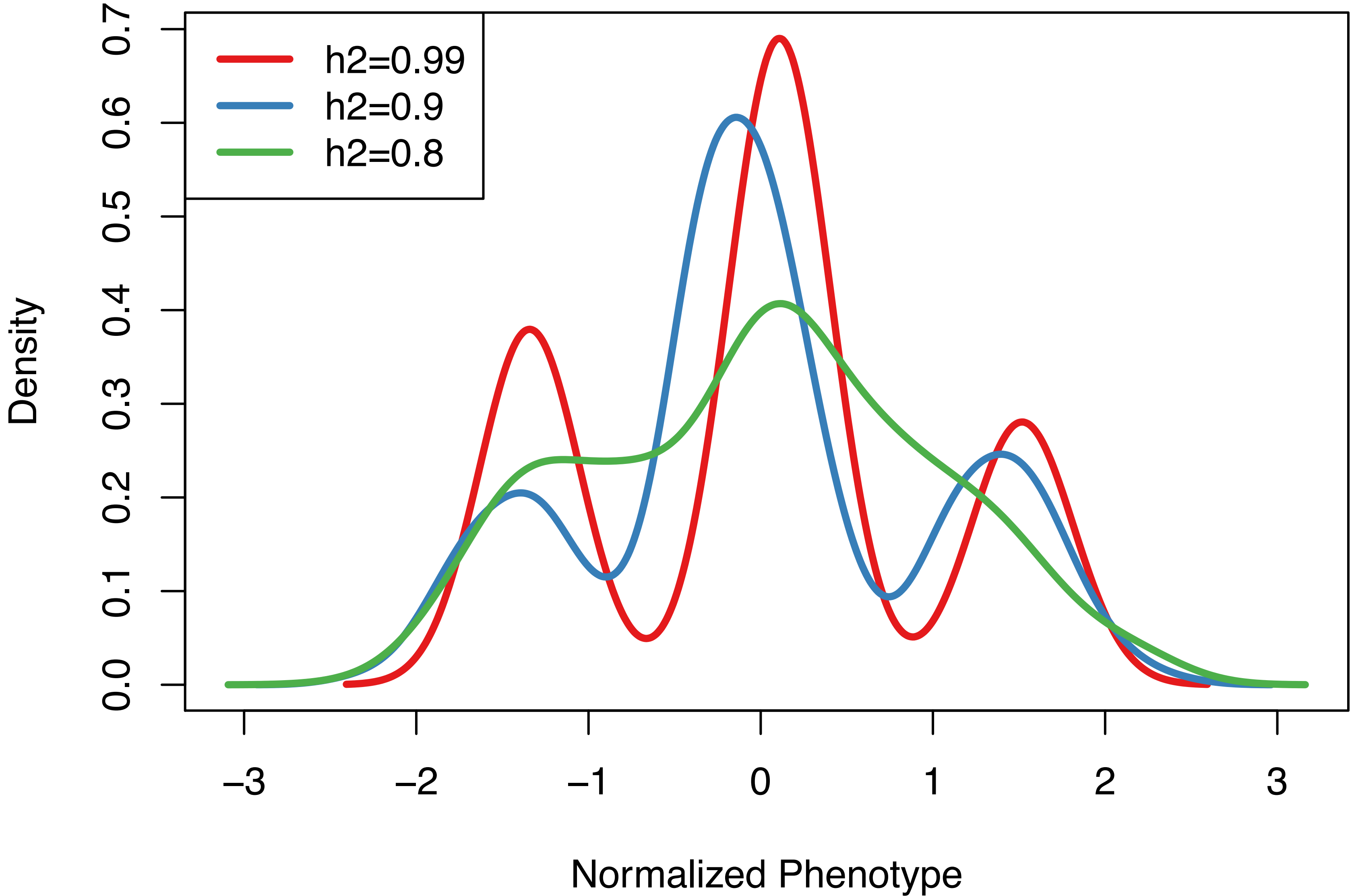
How do we simulate phenotypes?

- Simplest model:
  - There is 1 causal SNP.
  - Reference allele has no effect, but alternate allele has “some non-zero effect size”.



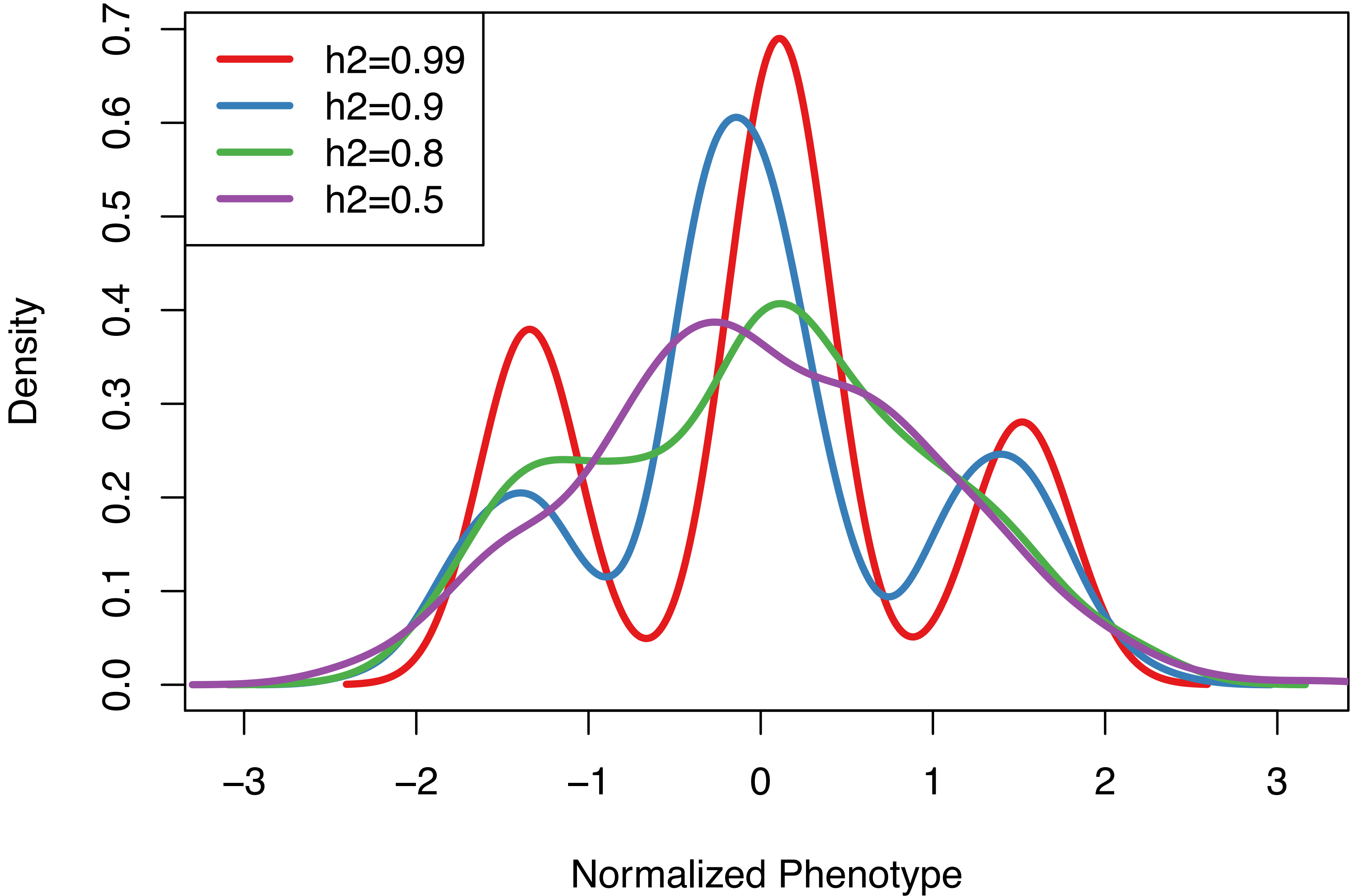
How do we simulate phenotypes?

- Simplest model:
  - There is 1 causal SNP.
  - Reference allele has no effect, but alternate allele has “some non-zero effect size”.



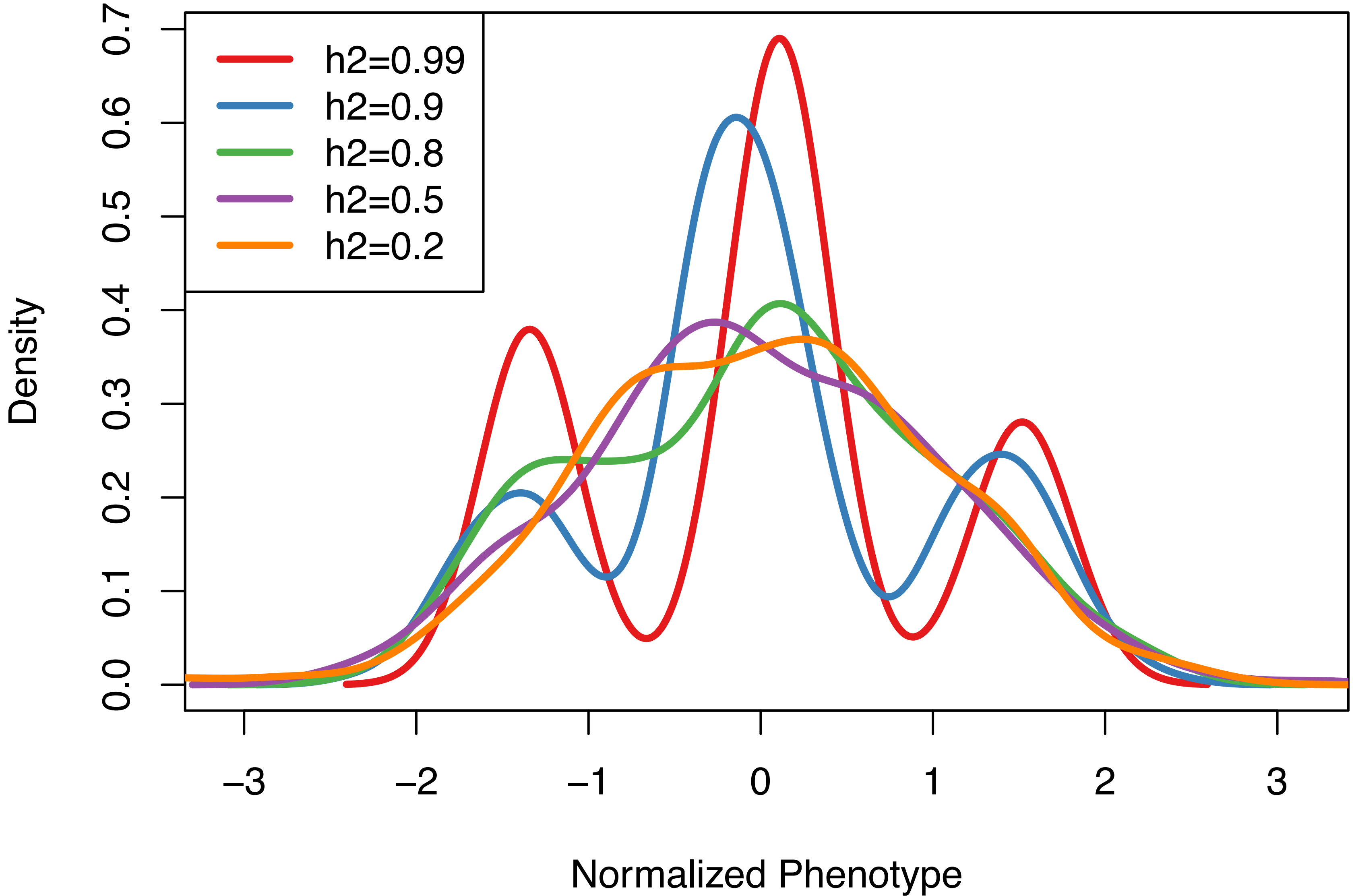
How do we simulate phenotypes?

- Simplest model:
  - There is 1 causal SNP.
  - Reference allele has no effect, but alternate allele has “some non-zero effect size”.



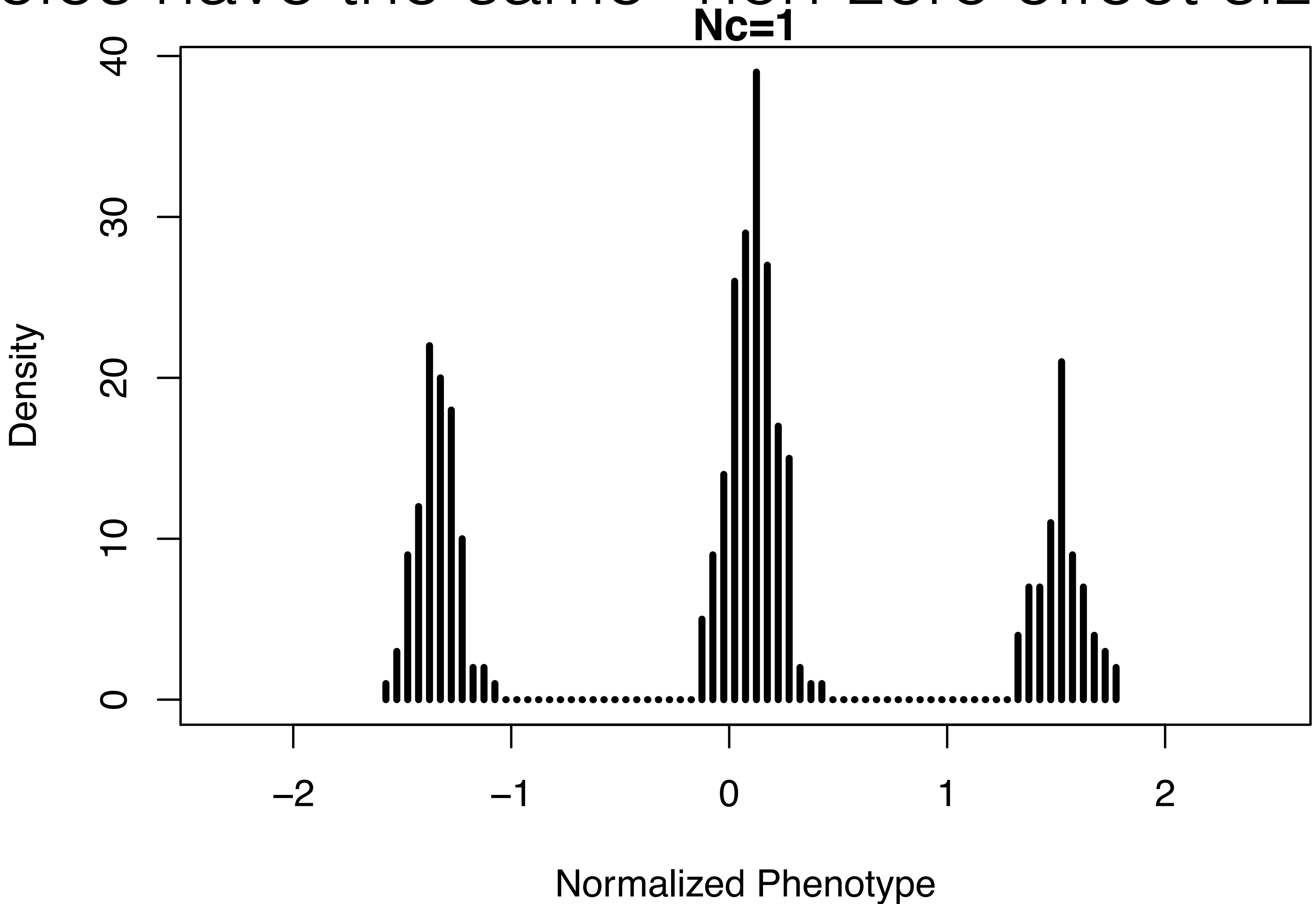
How do we simulate phenotypes?

- Simplest model:
  - There is 1 causal SNP.
  - Reference allele has no effect, but alternate allele has “some non-zero effect size”.



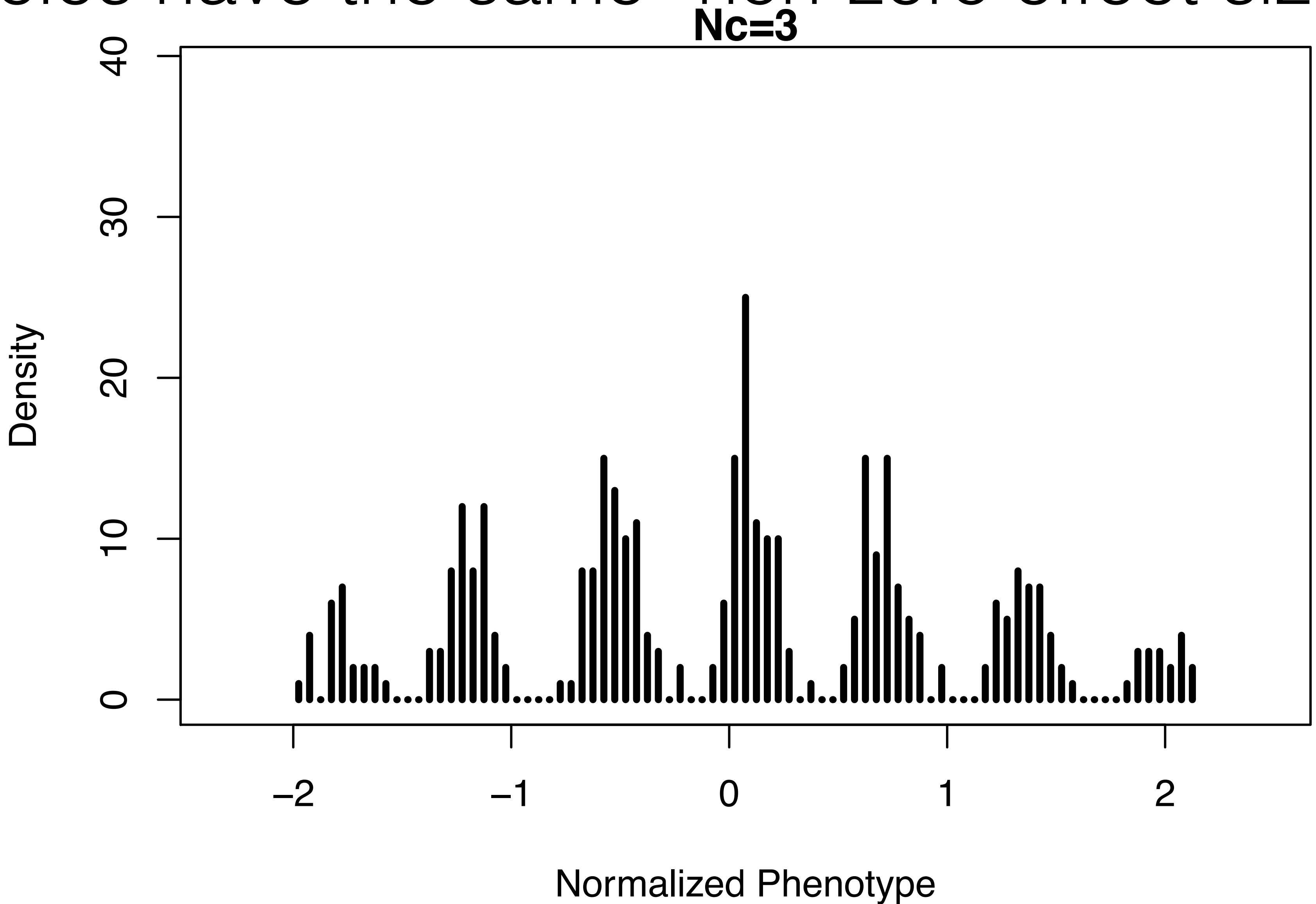
How do we simulate phenotypes?

- Less simple model:
  - There are **`Nc`** causal SNPs (h^2=0.99).
  - Reference alleles have no effect, but alternate alleles have the same “non-zero effect size”.



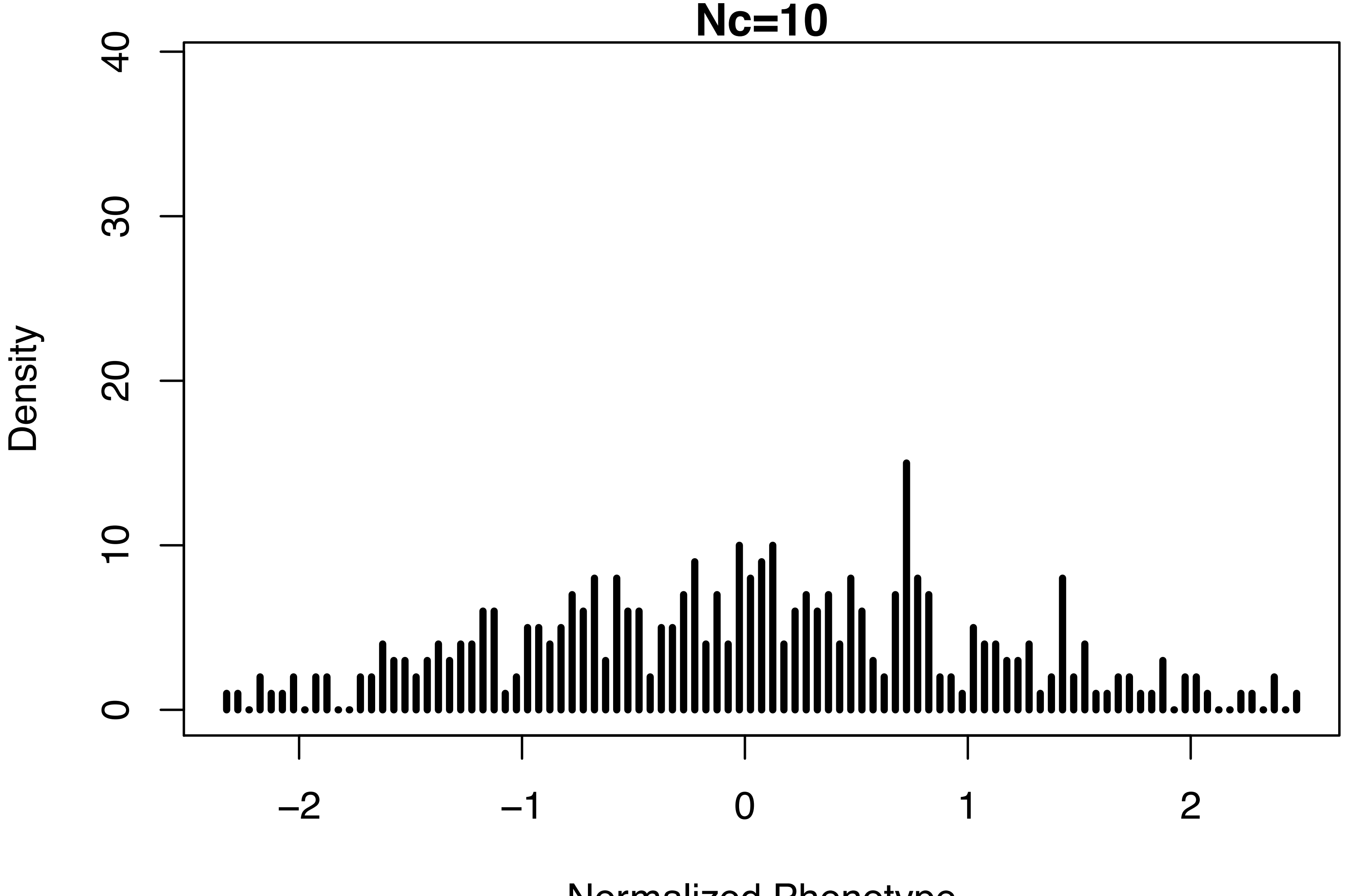
How do we simulate phenotypes?

- Less simple model:
  - There are **Nc** causal SNPs (h^2=0.99).
  - Reference alleles have no effect, but alternate alleles have the same “non-zero effect size”.



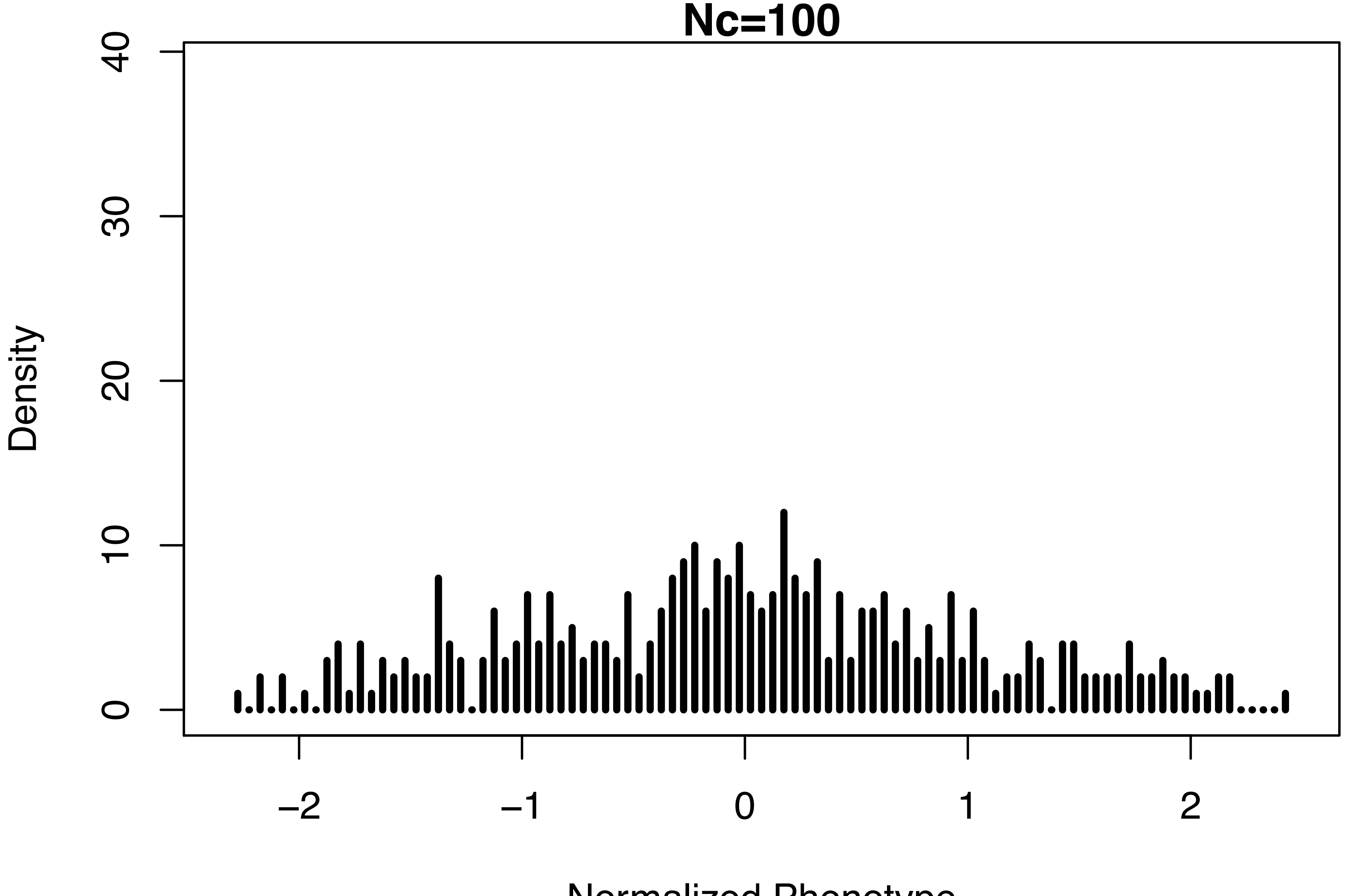
How do we simulate phenotypes?

- Less simple model:
  - There are **`Nc`** causal SNPs (h^2=0.99).
  - Reference alleles have no effect, but alternate alleles have the same “non-zero effect size”.



How do we simulate phenotypes?

- Less simple model:
  - There are `Nc` causal SNPs (h^2=0.99).
  - Reference alleles have no effect, but alternate alleles have the same "non-zero effect size".

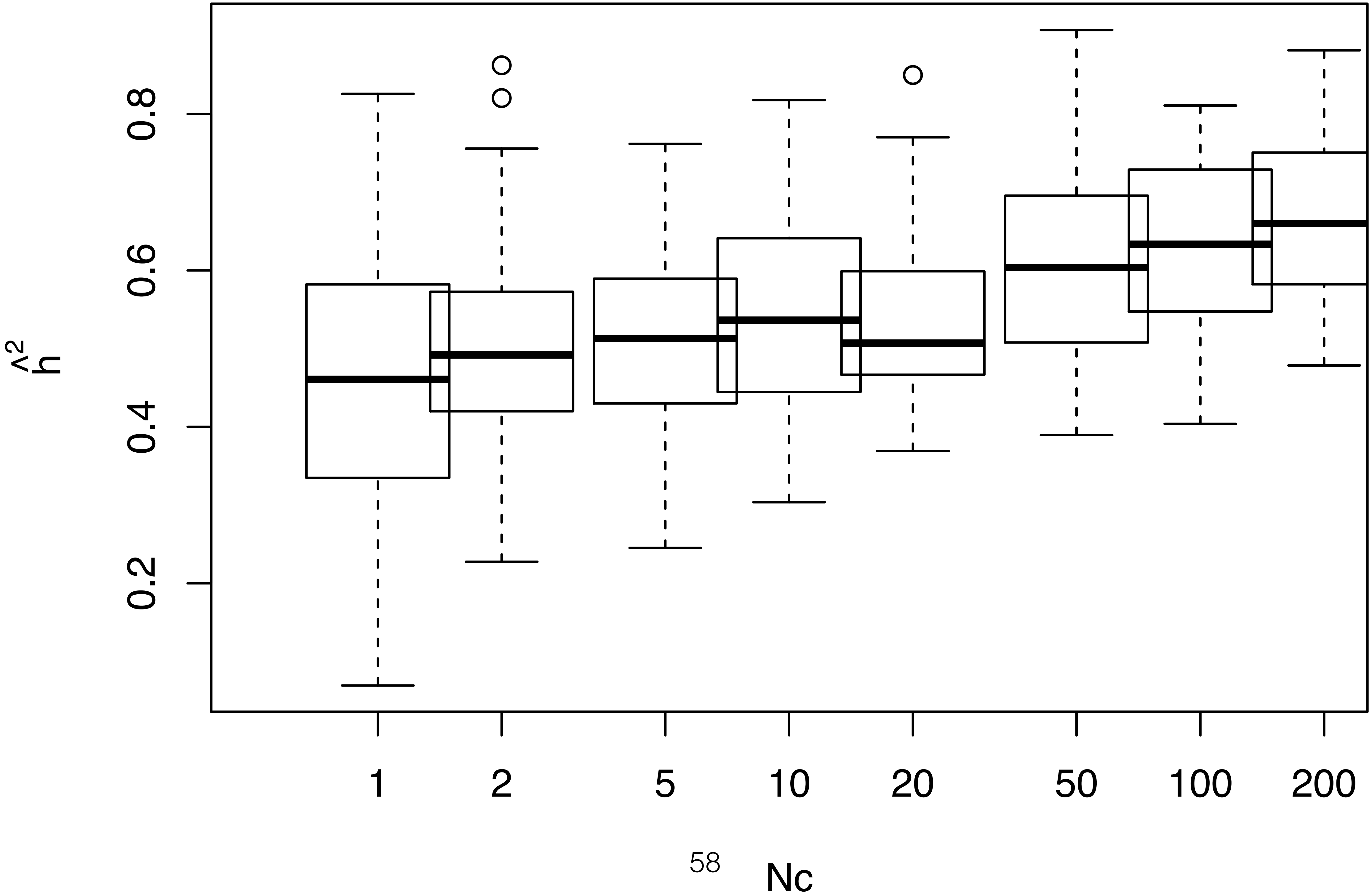


How do we simulate phenotypes?

- We are going to do the simulations in **`R`**!!
- Pick your favorite natural number (**`X`**).
- type:
  - **`Rscript  HEplay.R  Nc=X`**
- Who picked the smallest/largest number?
- Who got the smallest/largest mean(estimated h2)?

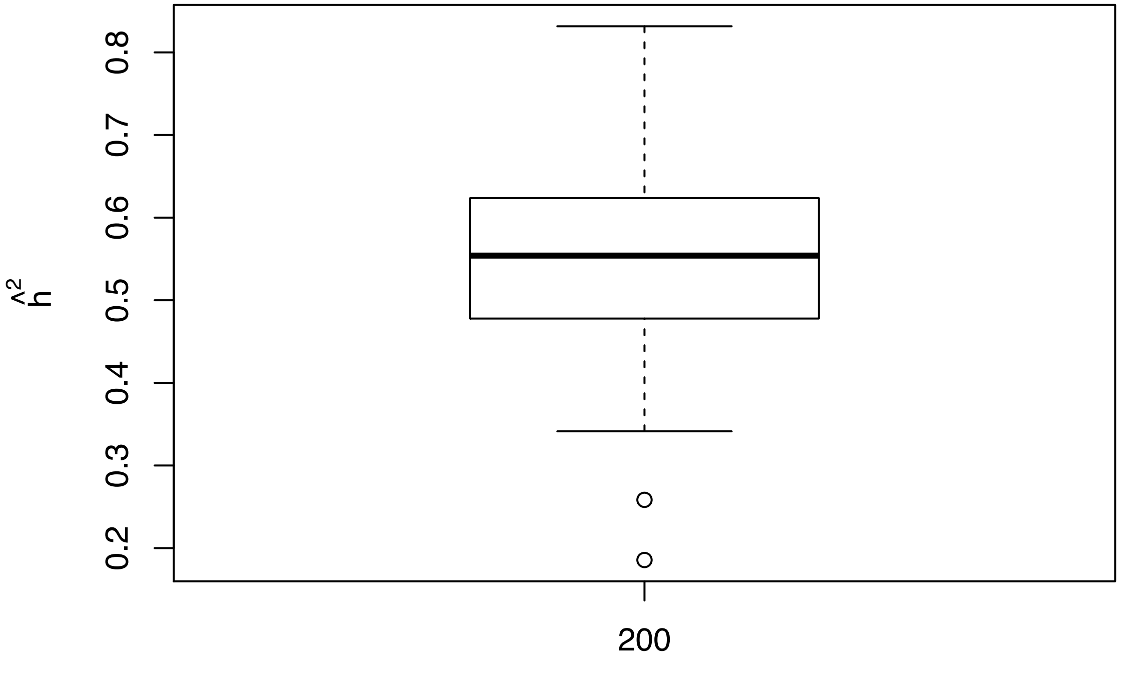
How do we simulate phenotypes?

- Are your results consistent with this?



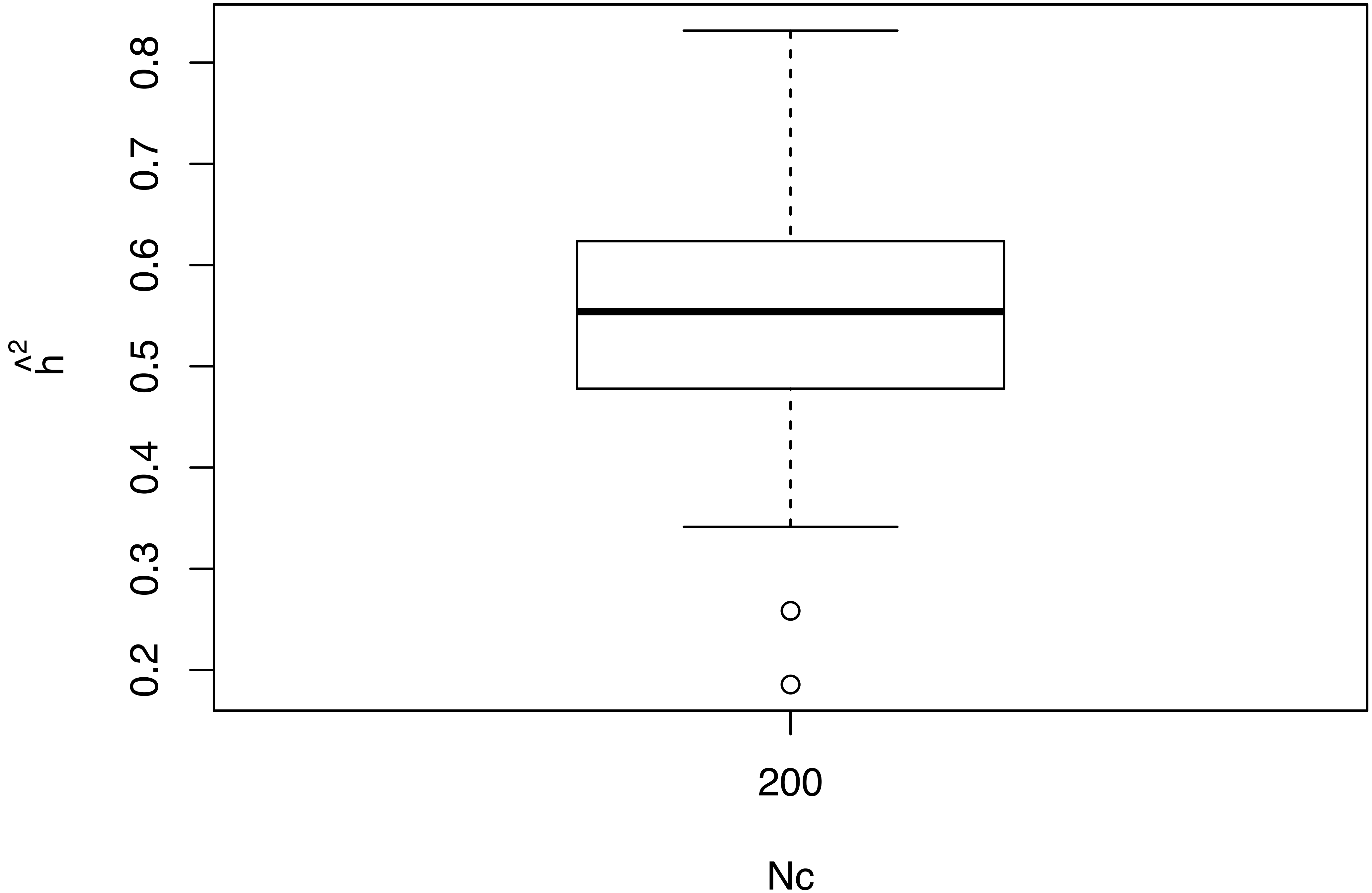
How do we simulate phenotypes?

- I've actually tricked you!
- By default, HEplay.R throws away all variants with MAF<0.05.
- You can change this in the simulation by typing:
  - `Rscript HEplay.R Nc=X minMAF=0`



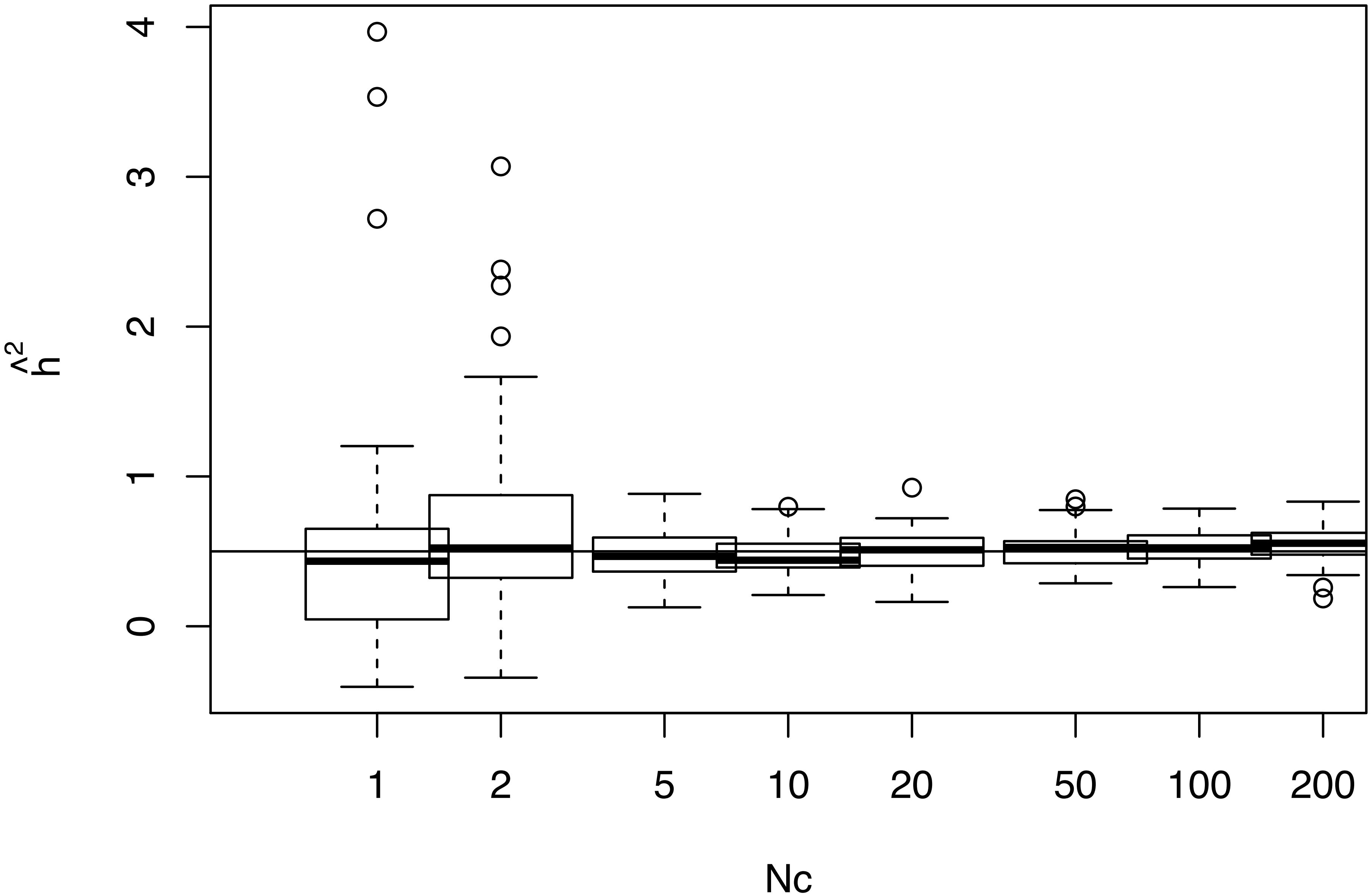
How do we simulate phenotypes?

- `Rscript  HEplay.R  Nc=X  minMAF=0`



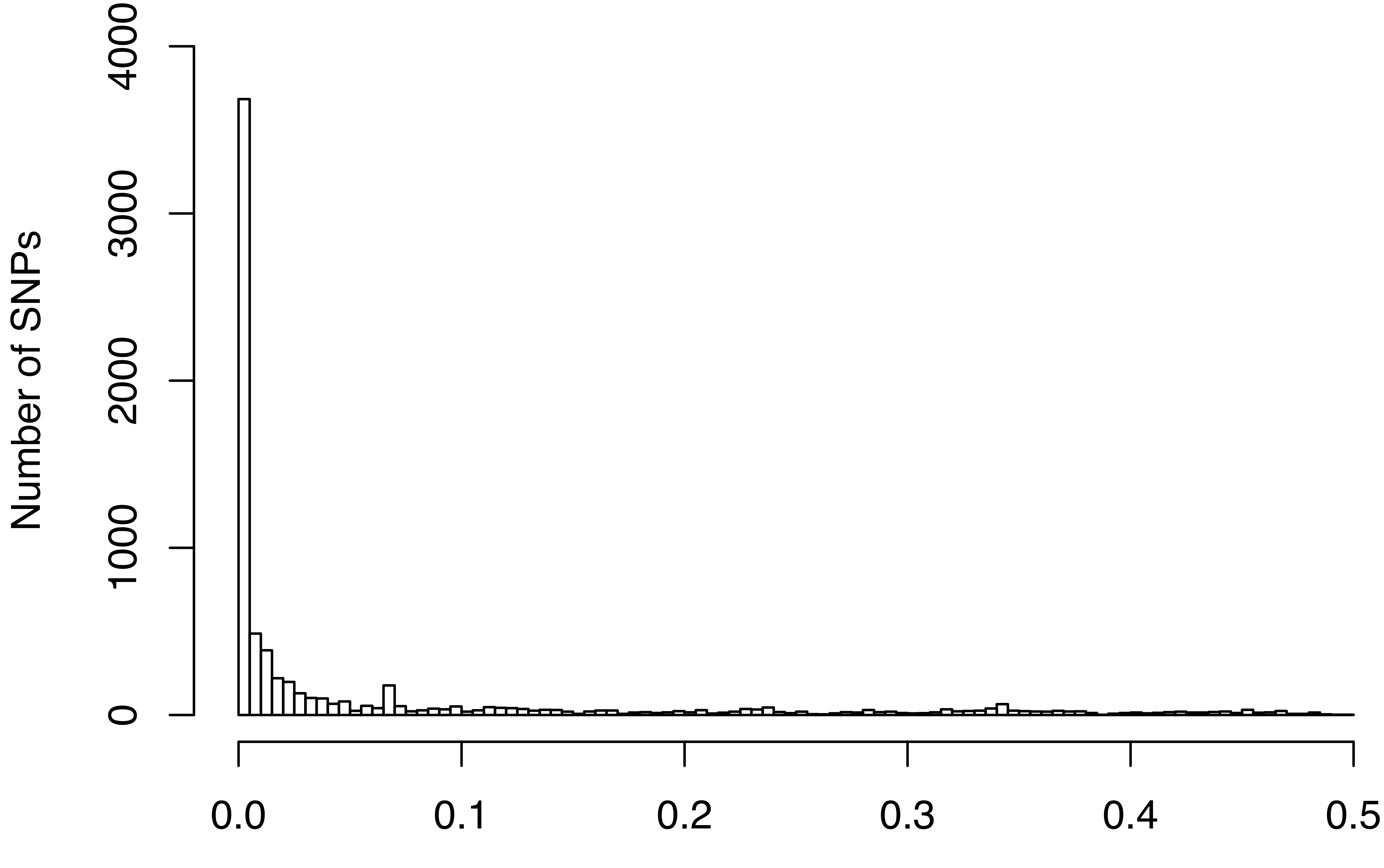
How do we simulate phenotypes?

- Who gets the largest/smallest estimate of h^2 now?
- There is another problem!!



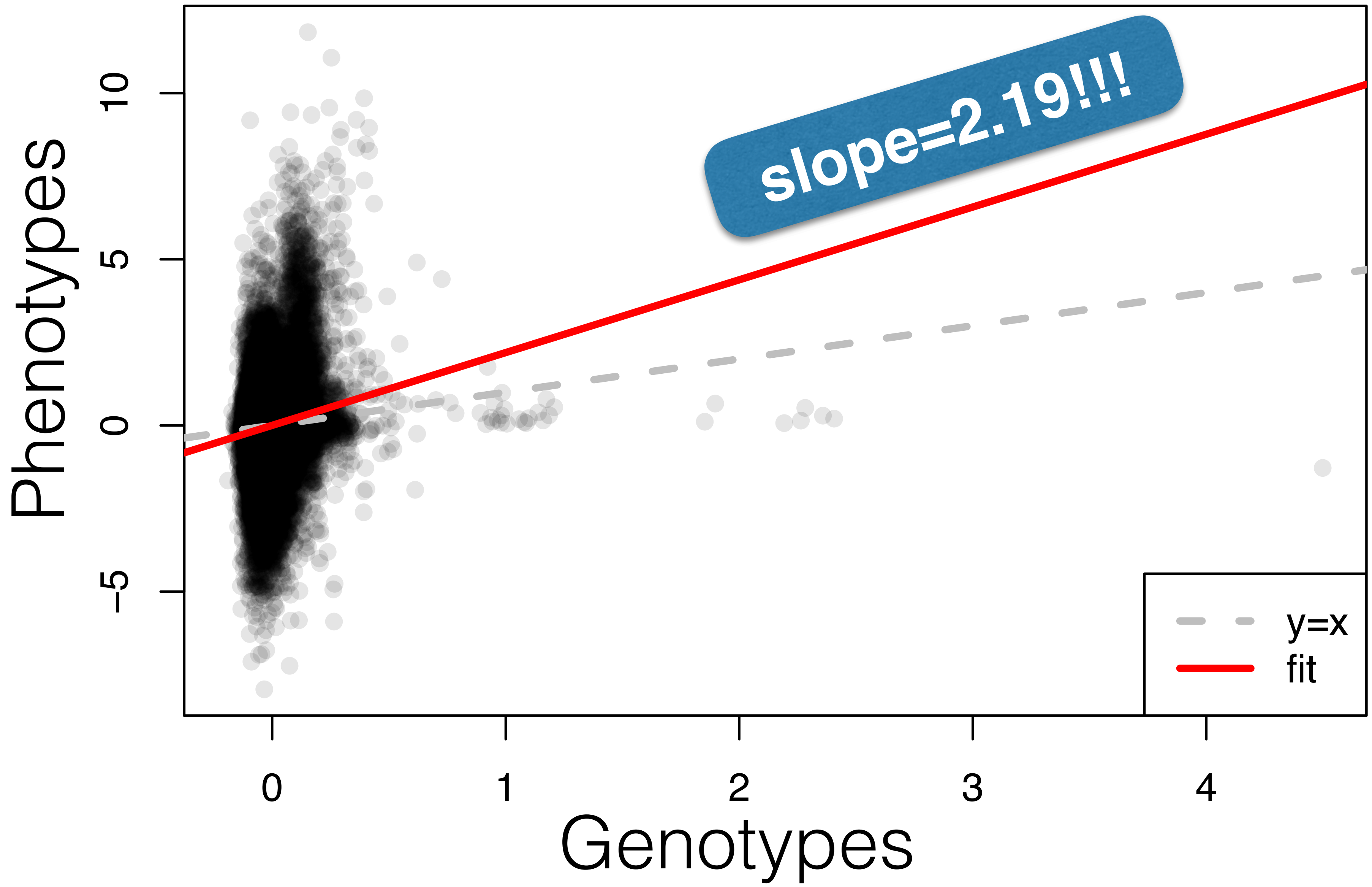
How do we simulate phenotypes?

- Here is the MAF distribution
- What happens if Nc=1 and your causal variant is a singleton?!



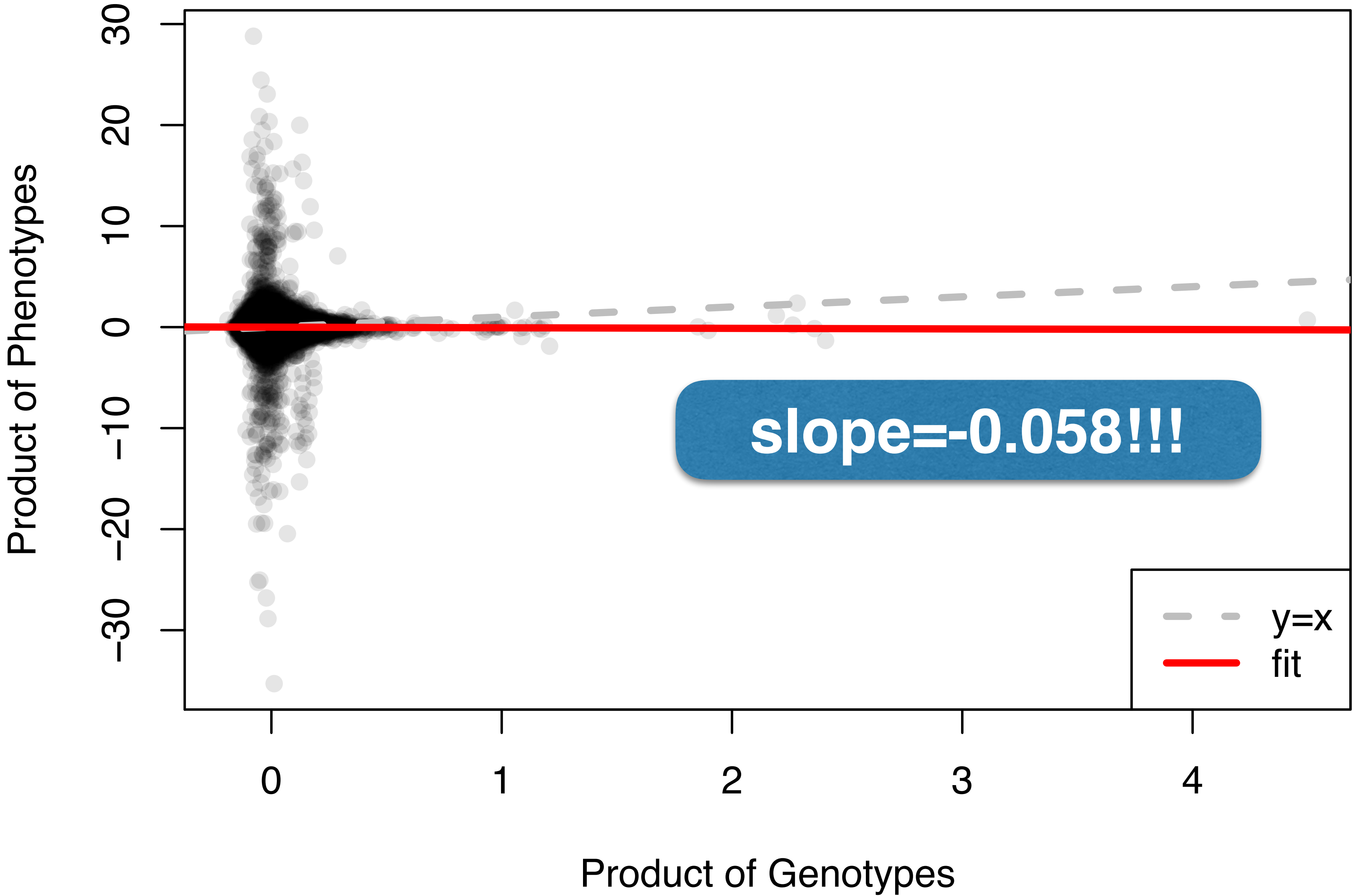
How do we simulate phenotypes?

- Here is the MAF distribution
- What happens if Nc=1 and your causal variant is a singleton?!



How do we simulate phenotypes?

- Here is the MAF distribution
- What happens if Nc=1 and your causal variant is a singleton?!



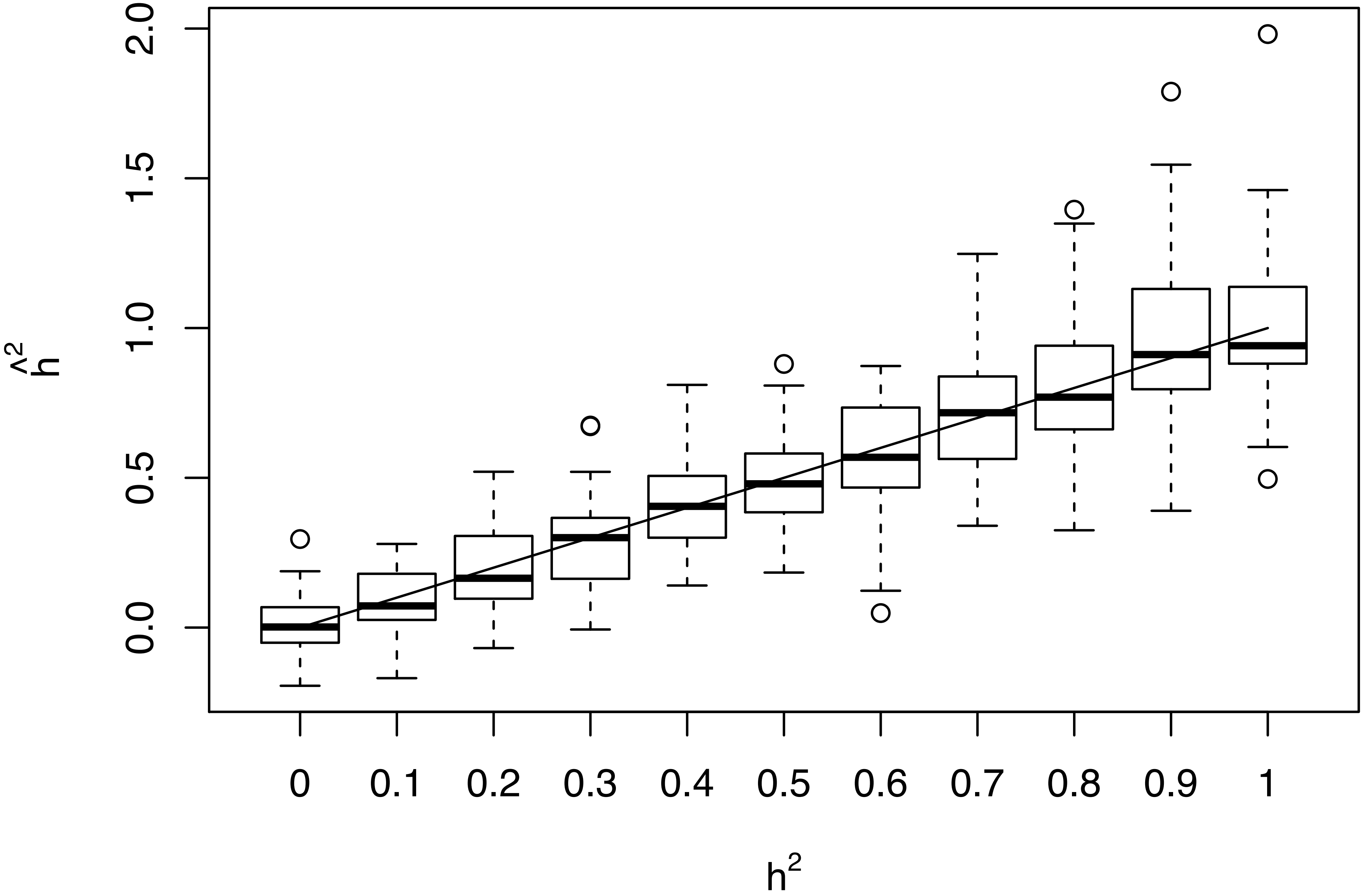
slope=-0.058!!!

How do we simulate phenotypes?

- Fortunately, most real traits are polygenic, so the algorithm works well.
- Pick your favorite value for h^2 between 0 and 1.
- Type:
  - `Rscript  HEplay.R  minMAF=0  h2=$h2`

How do we simulate phenotypes?

- Who picked the smallest/largest value of h2?
- Did you also get the smallest/largest mean(est(h2))?



How do we simulate phenotypes?

- Type:
  - `Rscript  HEplay.R  h`

```
rhernandez$ Rscript h2sim.R h
Options include:
   help (h)
        Prints out this help menu. For options below, default values are in parentheses.
   GENE=<gene>
        Note that only APOL1 is provided
   h2=<h2>
        Include any value of h2 between [0, 1]. (0.5)
   minMAF=<minMAF>
        Minimum MAF for variants included in analysis (i.e., exclude all variants with MAF < minMAF (0.05)
   INDIR=<input directory>
        Input directory for data file (.)
   Kbins=<K>
        The number of GRM bins you want to analyze (20)
   NSIMS=<NSIMS>
        Total number of simulations to run (25)
   CM=<CM>
        Model 1-3 for choosing causal variants (1)
   Nc=<Nc>
        The number of causal variants for analysis (10)
   K=<K>
        The number of frequency bins to choose causal variants from when CM=2 (1)
   fT=<fT>
        Frequency threshold for defining rare variants when CM=3 (0)
   …
```

How do we simulate phenotypes?

- There are many ways to play with this code to learn about how assumptions regarding the genetic model of a complex trait impact our inference of heritability.

- Type:

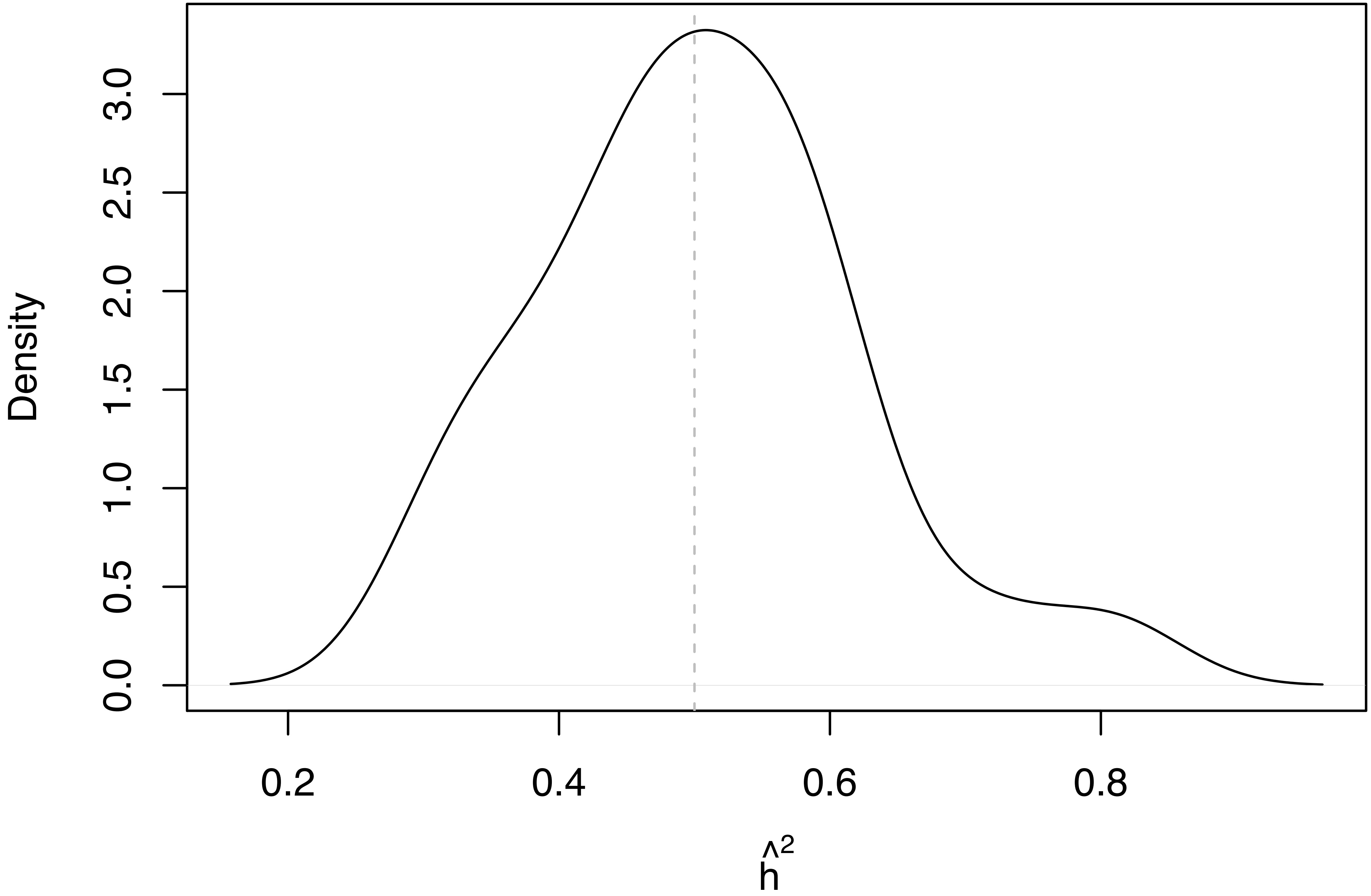
  - `Rscript   HEplay.R   PLOT=1`

  - This will create a file with the ugly name:

-
  ```
  ls
  h2hat_h2=0.5_minMAF=0.05_Kbins=20_CM=1_Nc=10_K=1_fT=0_fR=0_BM=1_fB=
  0.05.pdf
  ```

How do we simulate phenotypes?

- Once it is transferred, open it, and it should look something like this:

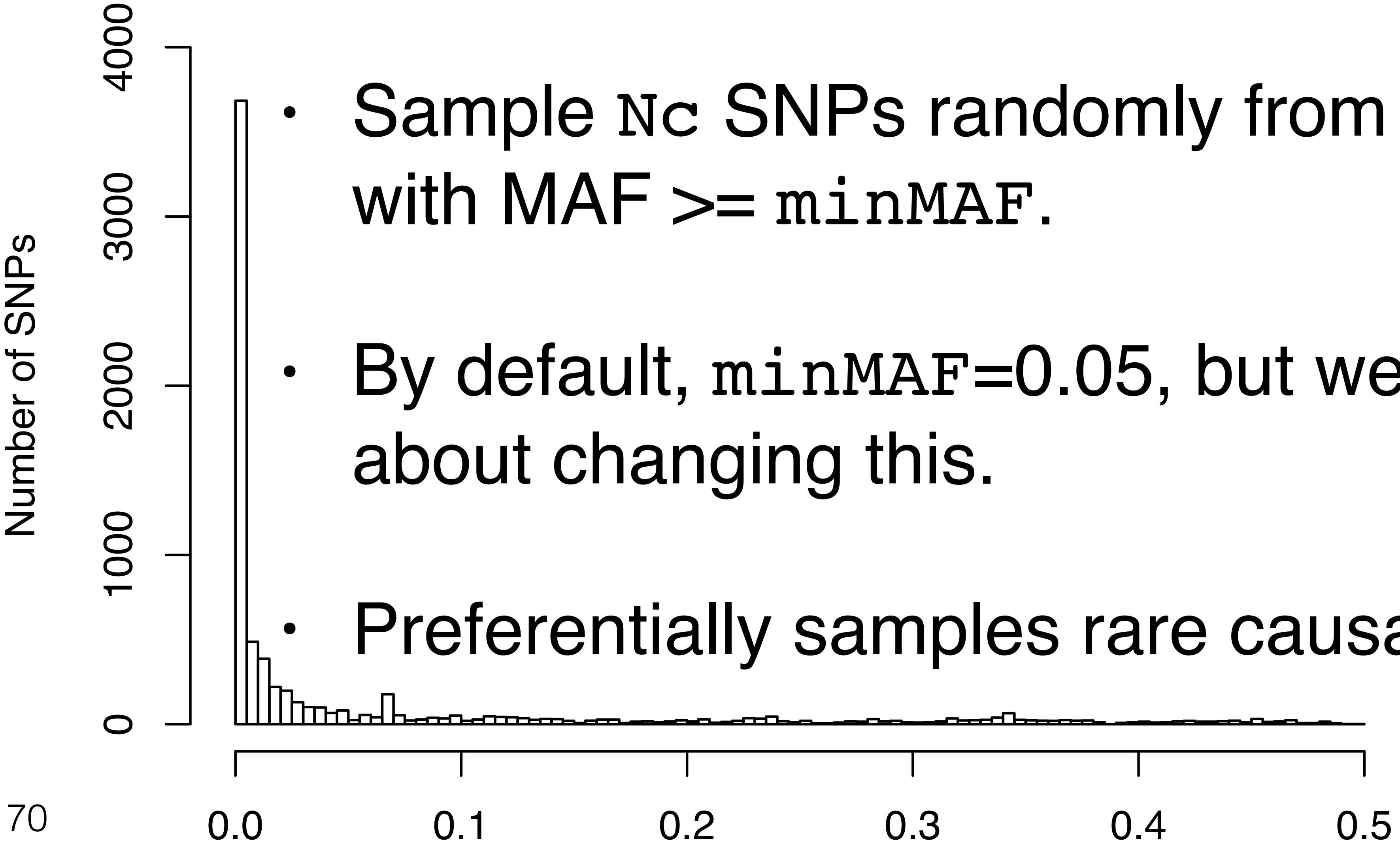


How do we simulate phenotypes?

- This script allows you to randomly drawn causal variants using 3 models using the option

  - `Rscript  HEplay.R  CM=<CM> [options]`

- So far, we have only used default: `CM=1`



  - Sample `Nc` SNPs randomly from the set of all SNPs with MAF >= `minMAF`.

  - By default, `minMAF`=0.05, but we already talked about changing this.

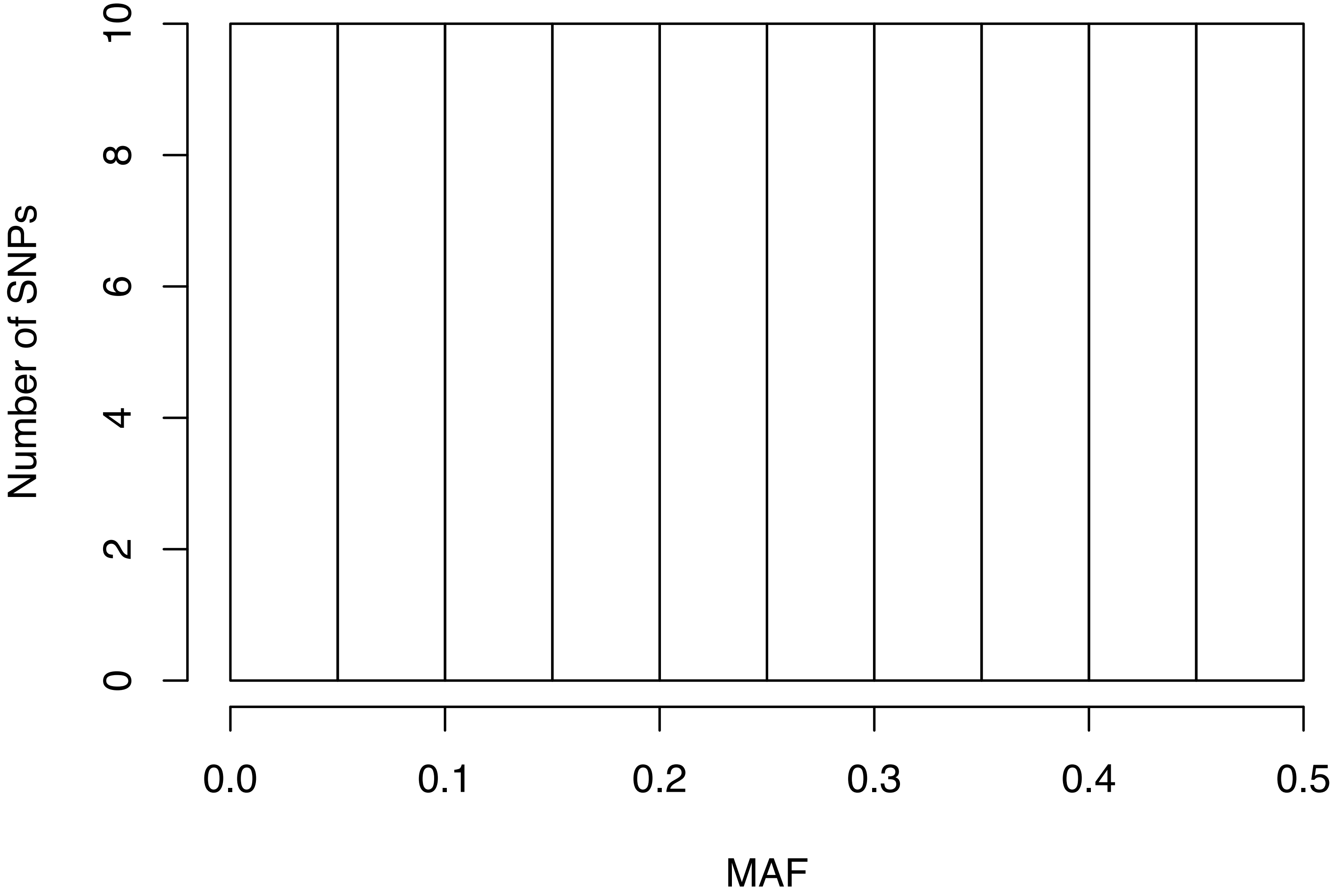
  - Preferentially samples rare causal variants.

How do we simulate phenotypes?

- This script allows you to randomly drawn causal variants using 3 models using the option
    - `Rscript  HEplay.R CM=<CM> [options]`
- `CM=2 K=<K>`
    - Randomly samples causal SNPs from `K` different bins.
    - For example, `K=2` would choose `Nc`/2 SNPs from (0,0.25) and `Nc`/2 SNPs from (0.25, 0.5)

How do we simulate phenotypes?

- This script allows you to randomly drawn causal variants using 3 models using the option
  - `Rscript  HEplay.R CM=<CM> [options]`
- `CM=2 K=10 Nc=100`



How do we simulate phenotypes?

- This script allows you to randomly drawn causal variants using 3 models using the option
  - `Rscript  HEplay.R  CM=<CM> [options]`
- `CM=3 fT=<fT> fR=<fR>`
  - Specify a frequency threshold (`fT`) for defining a rare causal variant.
  - Specify the fraction of causal variants that are rare (`fR`).
  - What would this do?
    - `CM=3 fT=0.01 fR=0.5`

How do we simulate phenotypes?

- Now that we have sampled causal variants, we need to specify their effect size.

- There are two ways to do this, specified using:

- `BM=<BM> [options]`

- So far, we have been using the default:

  - `BM=1 fB=0.05`

- This sets all effect sizes to be the same at 5%.

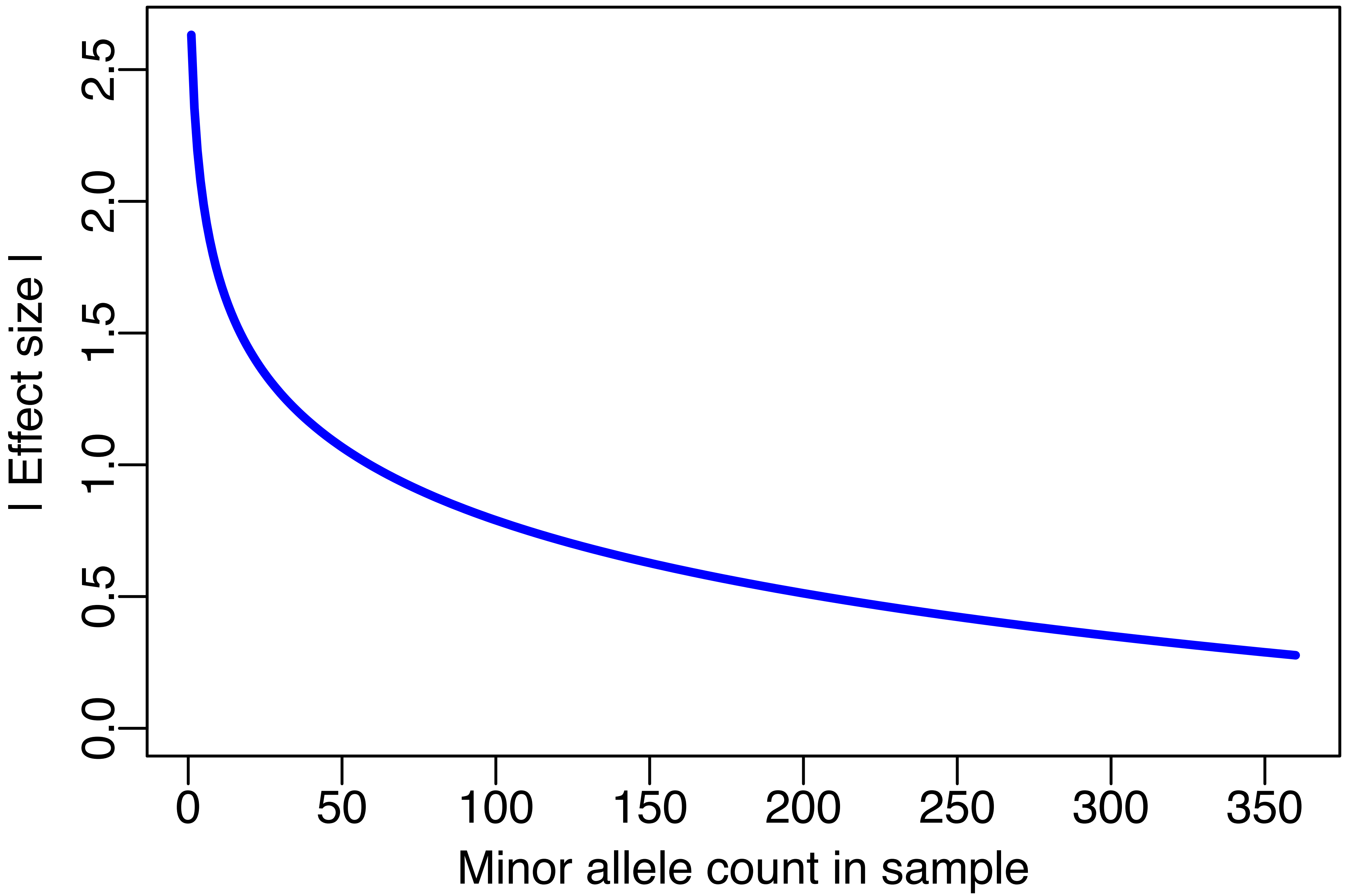
  - That is, each additional causal variant than an individual carries increases their phenotype by 5%

How do we simulate phenotypes?

- Now that we have sampled causal variants, we need to specify their effect size.
- There are two ways to do this, specified using:
- `BM=<BM> [options]`
- We can also make effect size a function of allele frequency:
  - `BM=2`
- A causal allele with frequency `x` will have effect size:
  - `-0.4*log`$_{10}$`(x)`

How do we simulate phenotypes?

- Under `BM=2` model, the effect size function looks like this:



Play!

- You can now combine these options to create interesting genetic models of complex traits!

- Here are a couple of examples (what do they do?):

    -
      ```
      Rscript HEplay.R CM=2 K=10 minMAF=0 BM=2
      ```

    -
      ```
      Rscript HEplay.R CM=3 fT=0.01 fR=0.9 minMAF=0
      ```

    -
      ```
      Rscript HEplay.R CM=3 fT=0.01 fR=0.9 minMAF=0 BM=2
      ```

- Whoever creates the model with the largest bias wins!