Simulating phenotypes using coalescent data

Mark Reppell

July 21, 2015

Putting the coalescent to work

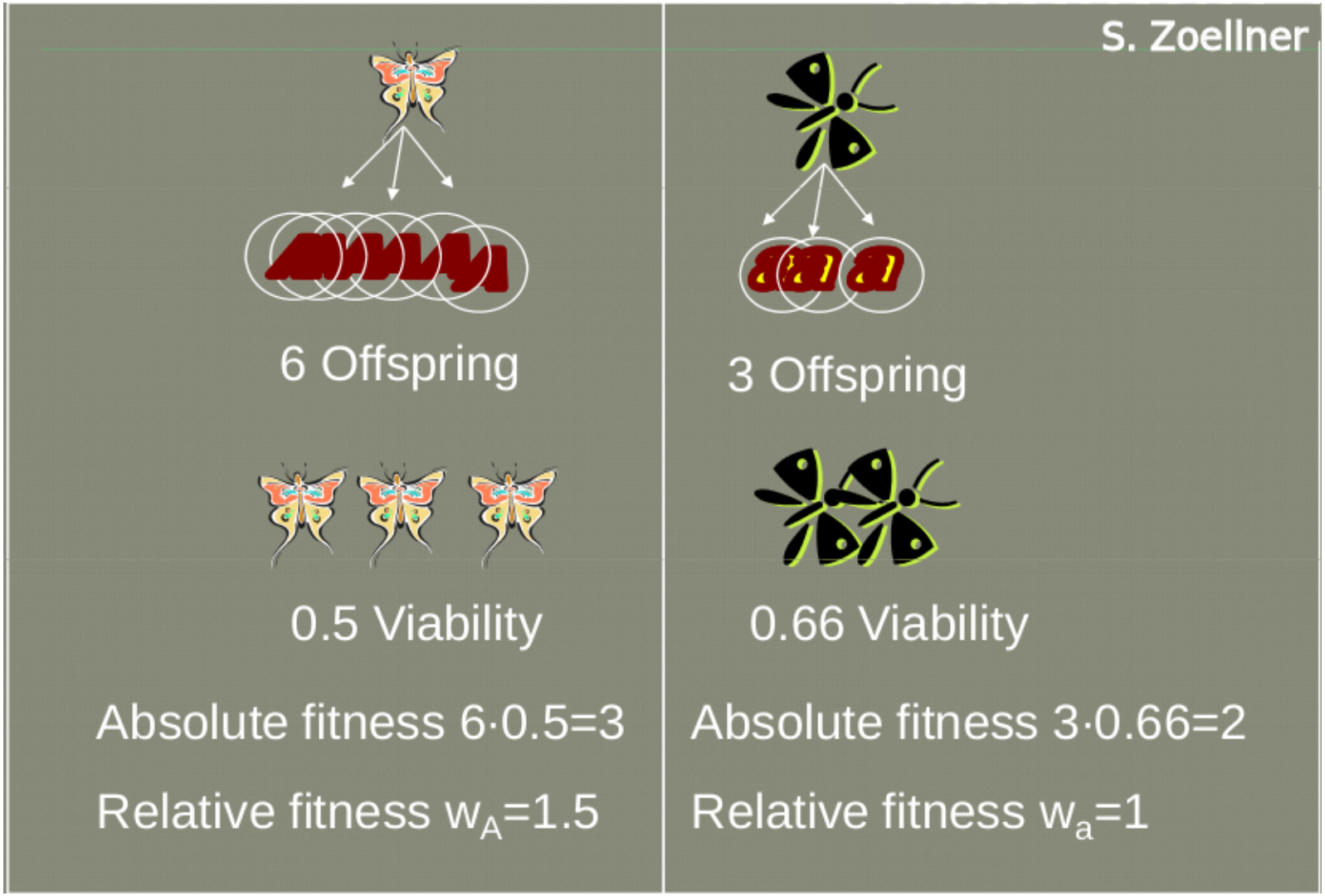
Outline

A tangent on natural selection

Methods of simulating phenotypes

- Inheritance models and working with multiple casual loci
- Quantitative trait models
- Binary phenotype models

A function of both viability and fertility



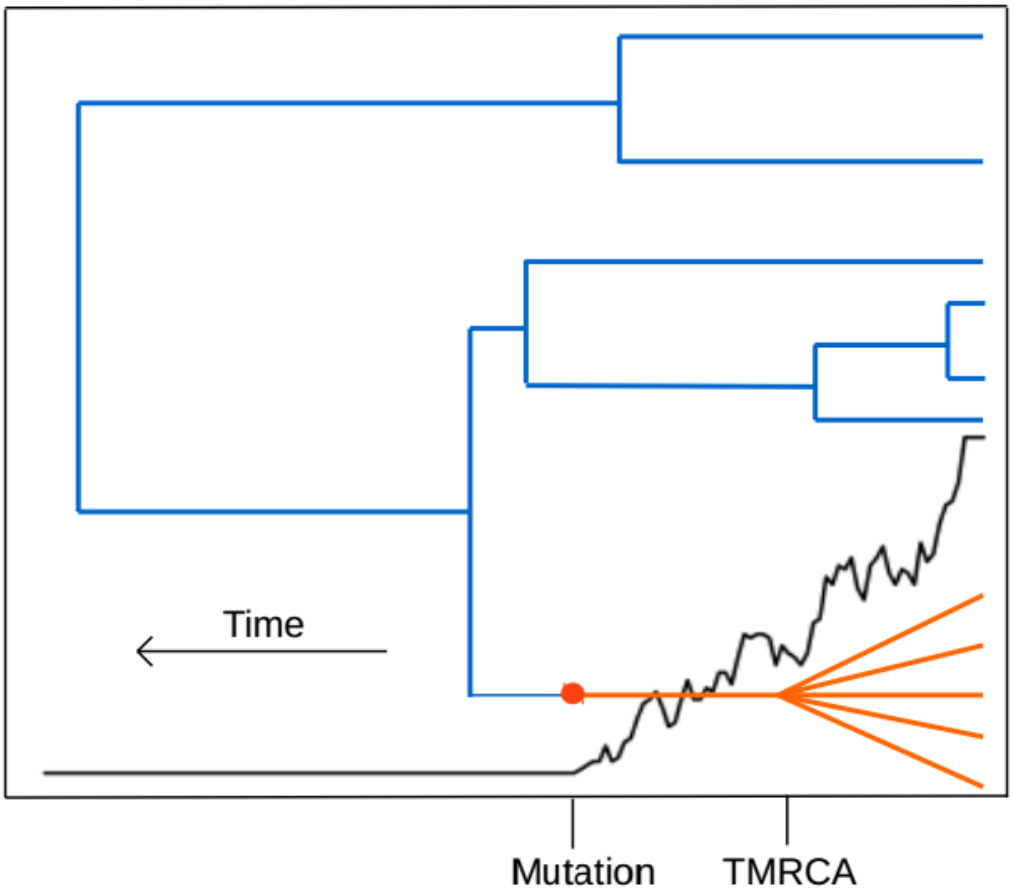
Natural Selection and the Wright-Fisher model

In the Wright-Fisher model we assume every chromosome in the preceding generation is equally likely to be ancestral to current generation chromosomes.

Natural selection violates this assumption, individuals with higher fitness are more likely to be ancestral than those of lower fitness

No convenient exponential process developed, requires a more complicated approach

Modeling positive selection at a single locus



Selection can be modeled as a kind of combination of population structure and growth, with the allele frequency shrinking backwards in time

Image from Joel Smith

Negative and multilocus selection

Complex selection scenarios force us to return to forward time simulations:

1) Set total number of sites, mutation model, and initial population size and structure

2) For each diploid individual calculate relative fitness based on current variant load

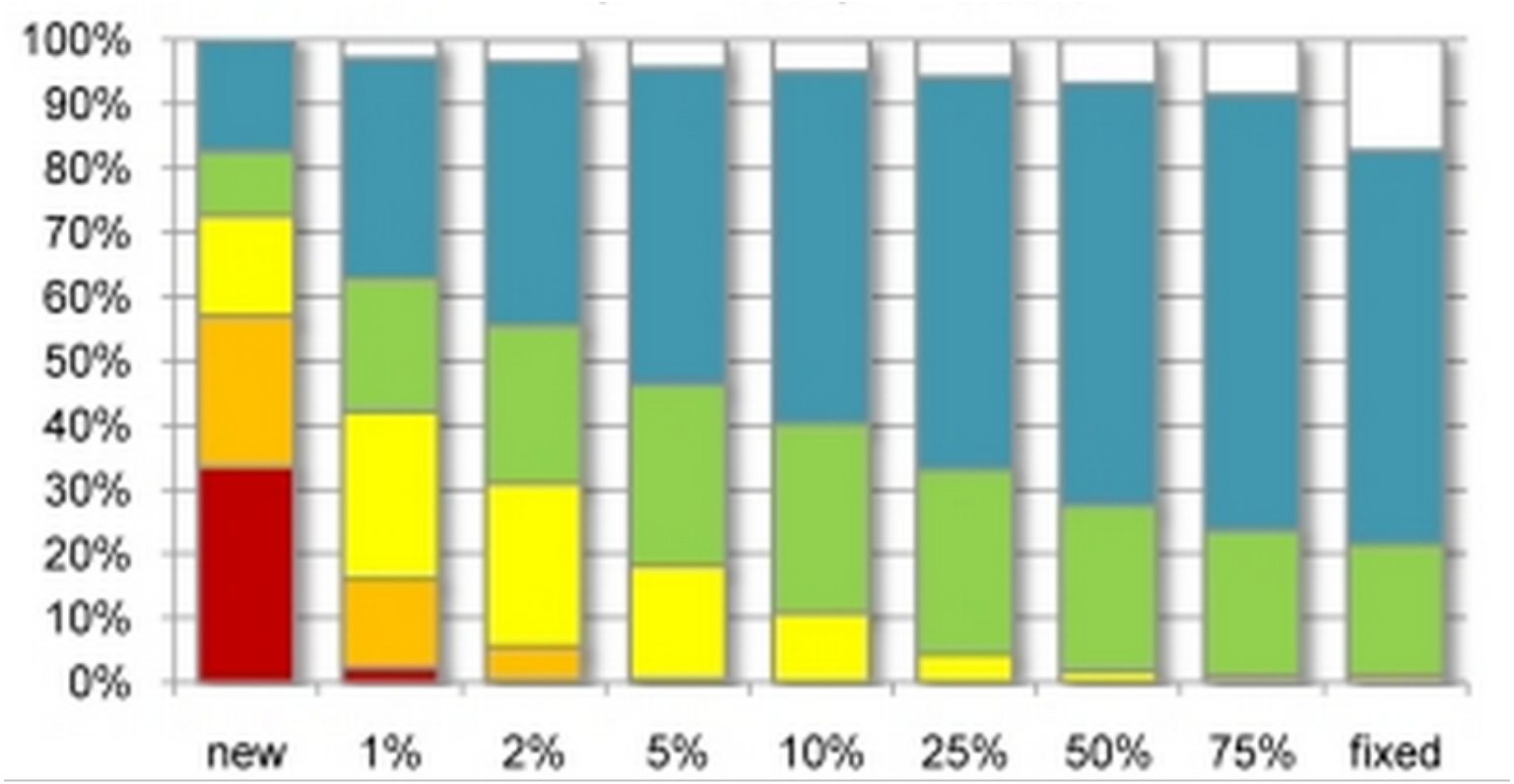
3) For each individual in next generation (possibly different number than previously) draw parental chromosomes based on fitness

4) Migration and recombination events if applicable

5) Mutation events, choose sites, draw fitness effect from user specified random distribution

6) Repeat for desired number of generations

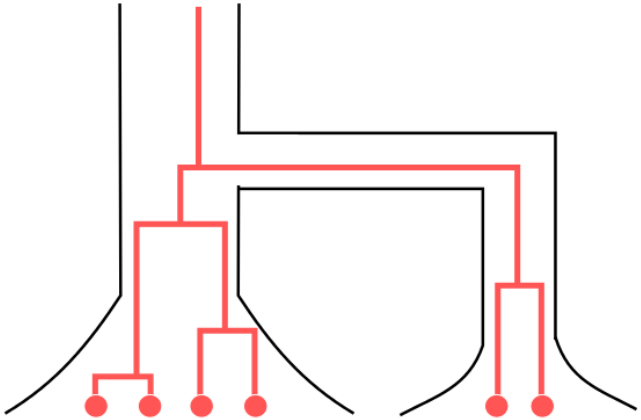
Inferred selection effects in human populations



Fitness effects at different allele frequencies inferred in 100 chromosomes using African American demographic history

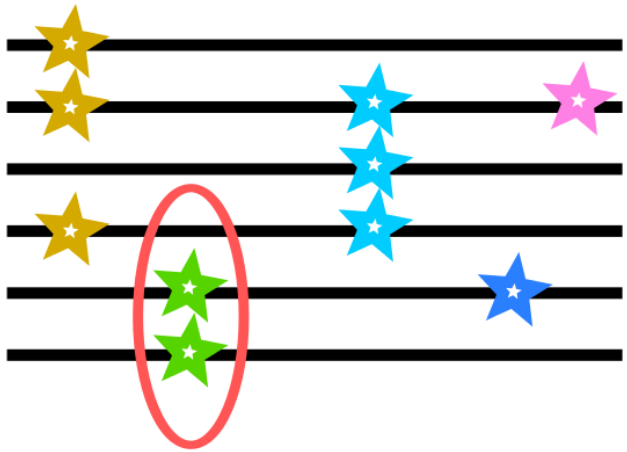
Boyko *et al.* 2008

Overview



Simulate Loci under desired demography

Choose causal variant(s)



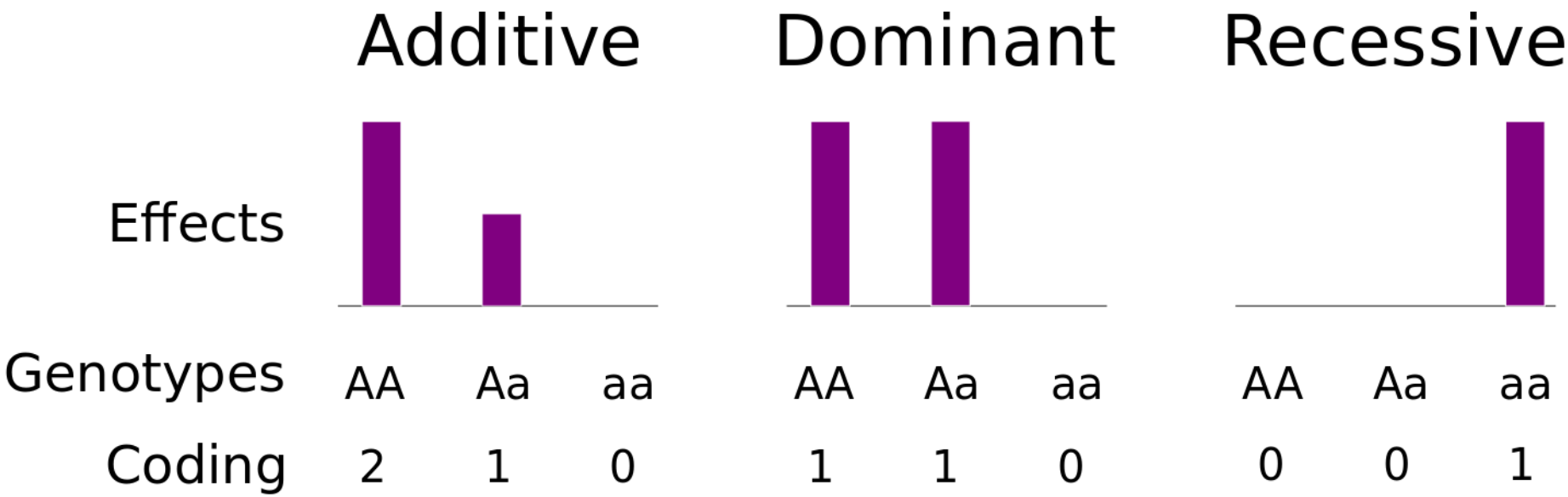
Inheritance Model

Sample	Pheno	Site1	Site2	...
1	1	1	0	
2	0	2	0	
3	0	0	1	
4	1	1	0	
...	...	...	...	

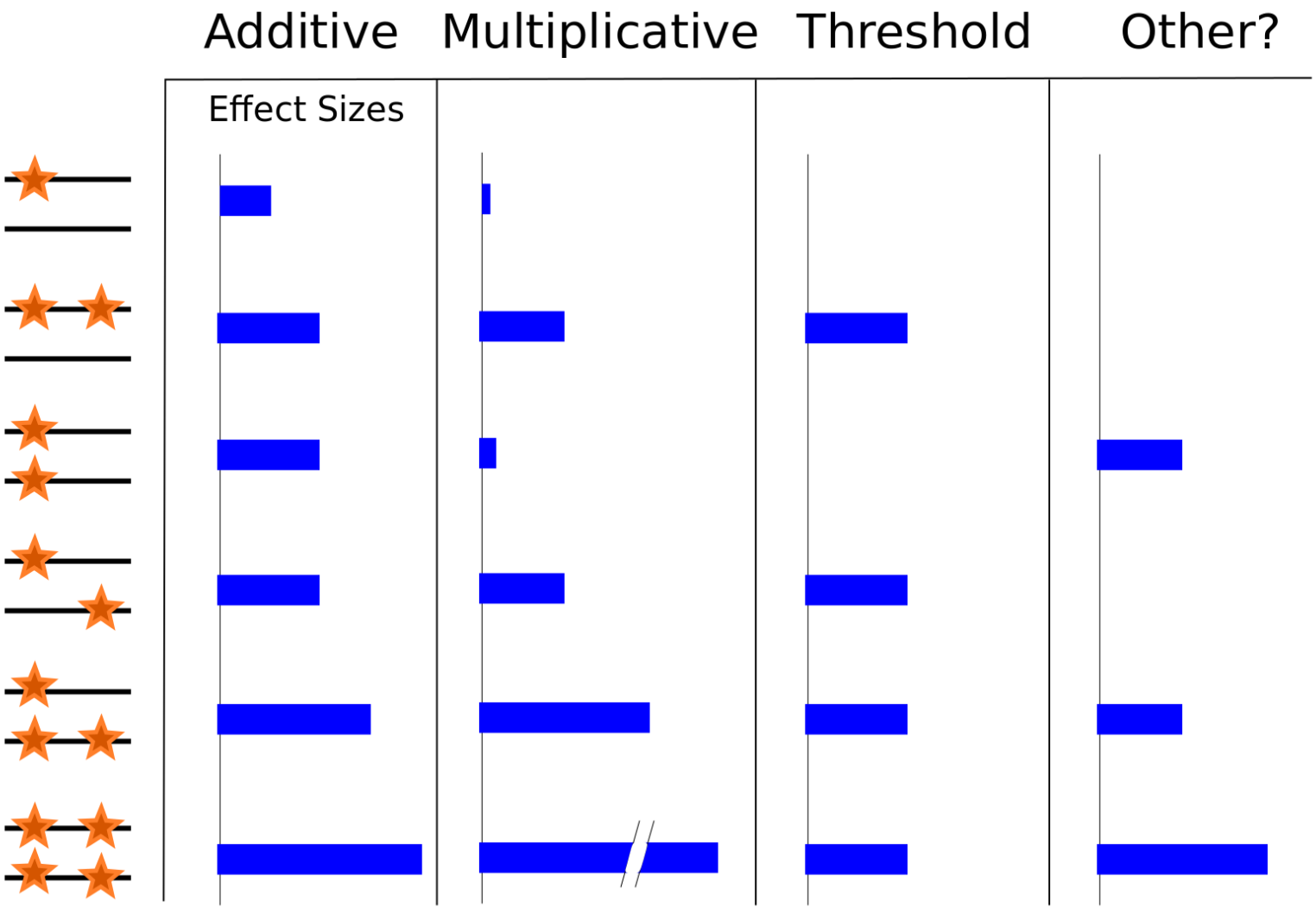
Generate Phenotypes

Data for Downstream Analysis

Biallelic inheritance models

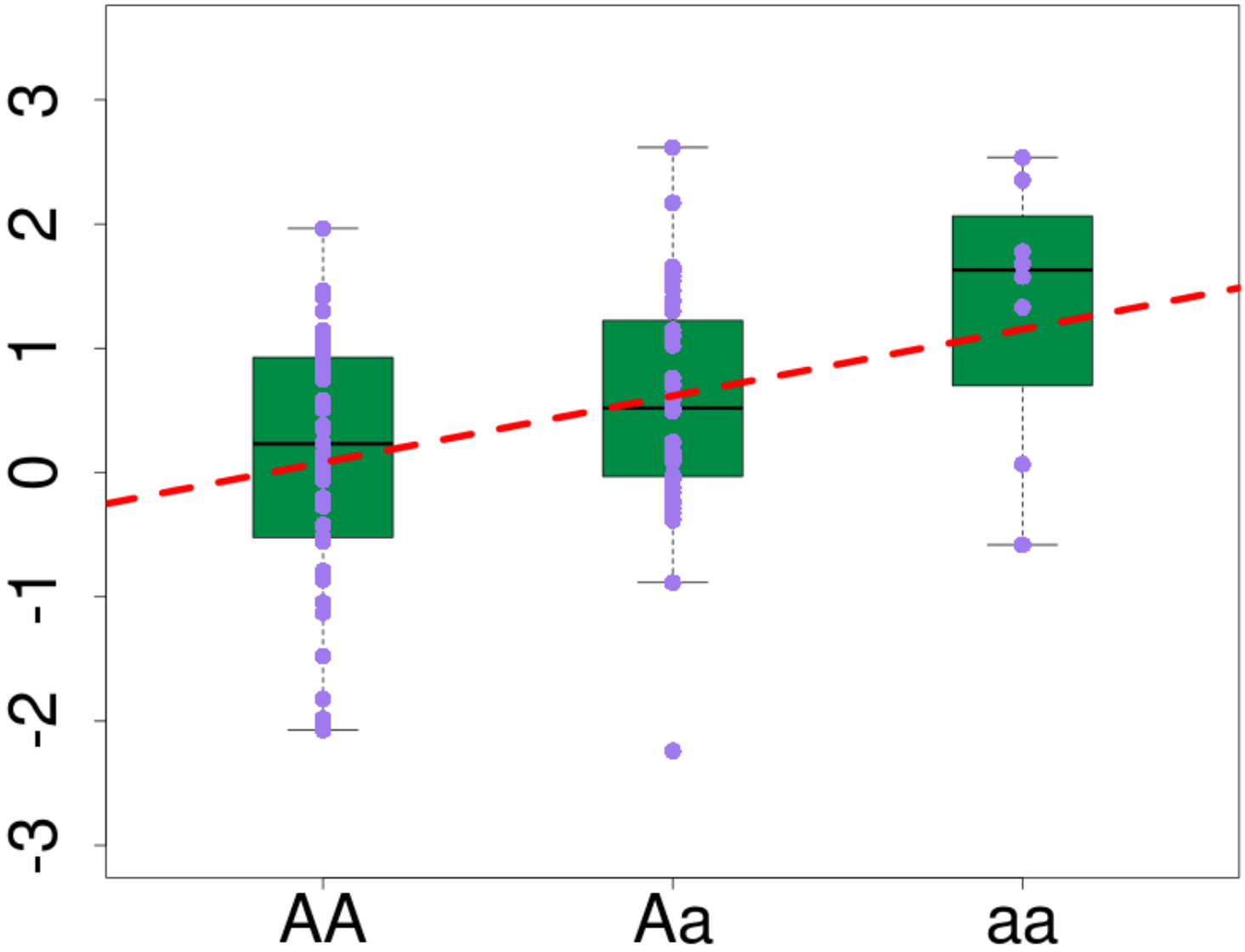


Multiple causal alleles - epistasis



The space of possible models is frighteningly large

Simulating Quantitative Traits



Variance explained QT model

If we assume causal loci are independent, to generate a QT with mean μ and variance σ^2 in a sample of size n, we can use the linear model

$$Y_i = \mu + \sum_{g=1}^{m} \beta_g X_{i,g} + \eta$$

Where Y_i is the QT value for individual $i \in (1, n)$, $X_{i,g}$ is the variant coding at site $g \in (1, m)$ for individual i, β_g is the effect size of variant g, and η captures the trait variability not explained by the genetic variants being modeled.

Variance explained QT model

If we wish the genetic factors to explain a proportion p of trait Y variability we can use the following

$$Var(Y) = \sum_{g=1}^{m} \beta_g^2 Var(X_g) + Var(\eta)$$

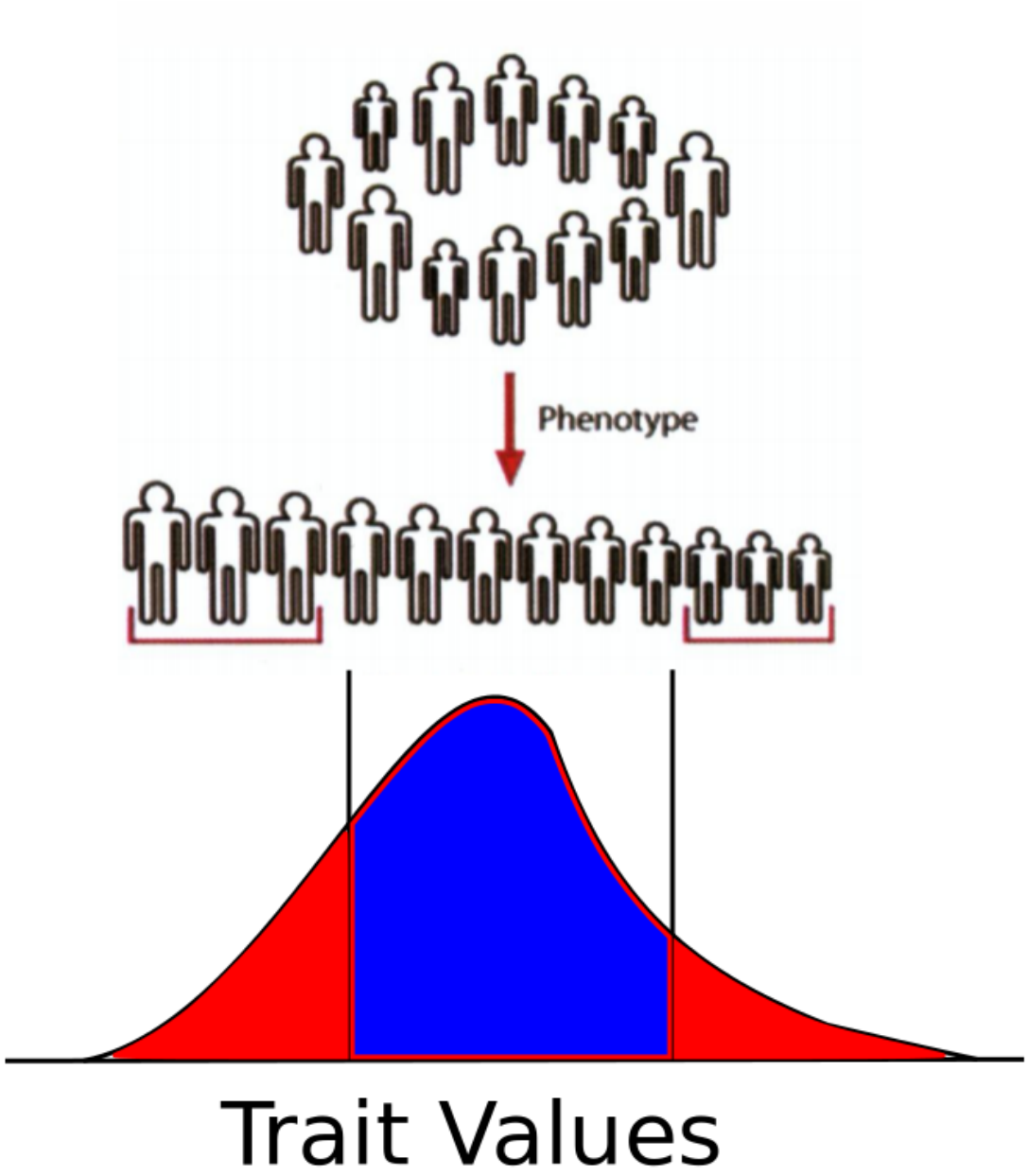
If we set $\eta \sim N(0, \sigma^2(1-p))$, we can calculate $Var(X_g)$ from the data or use $2\hat{p}(1-\hat{p})$ where $\hat{p}$ is the sample allele frequency, then any combination of β_g such that

$$p\sigma^2 = \sum_{g=1}^{m} \beta_g^2 Var(X_g)$$

Will explain the desired amount of trait variance.

In the simple case of a single causal variant we set $\beta = \sqrt{\frac{p}{Var(X)}}$

Extreme phenotypes models



Simulating extreme phenotype samples

Approach 1

Simulate random population with QT

Take upper/lower $p^t h$ percentile of QT values as extreme, keep only these samples

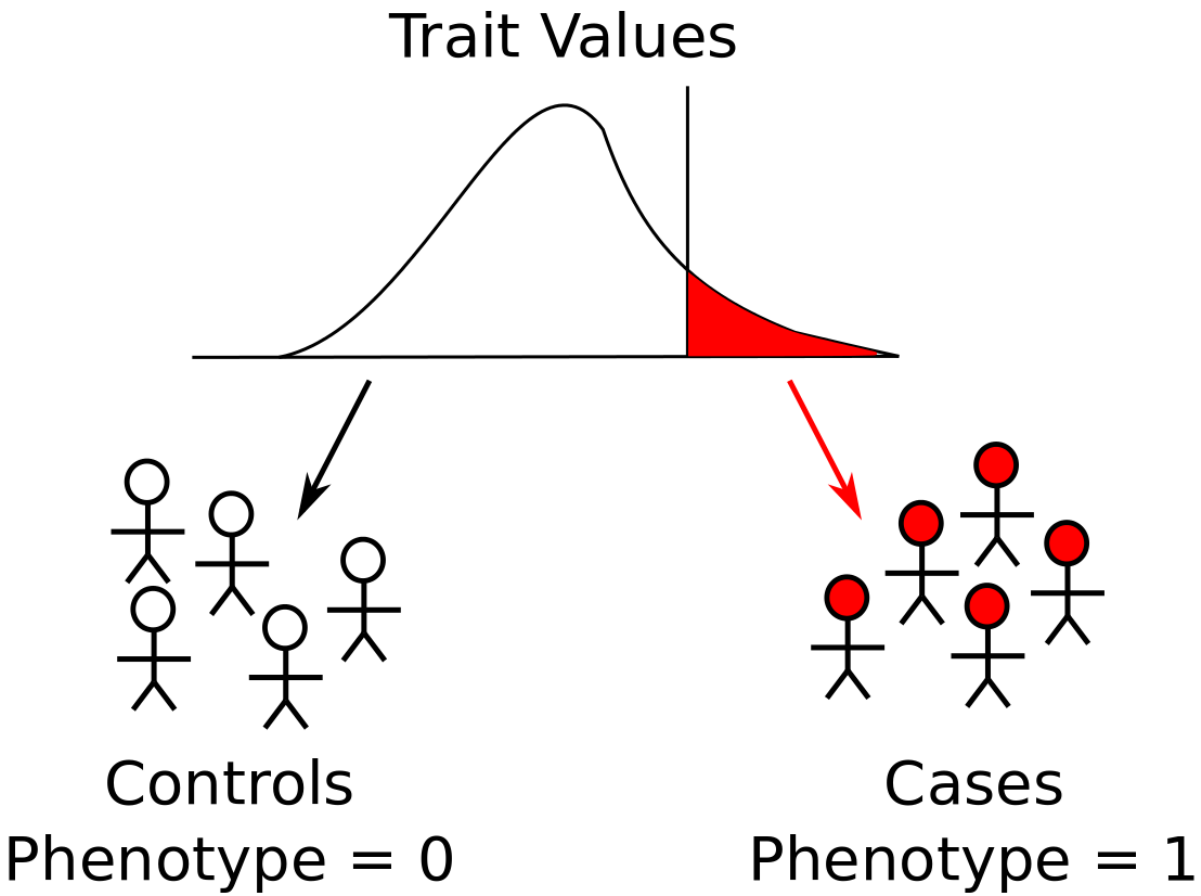
Approach 2

Define upper/lower QT value cutoffs

Repeatedly generate samples, keep if $|QT| >$ cutoff

Continue until achieve desired sample size

Naive approach to binary traits



Simulate QT using linear model, set threshold T, and if $QT > T$ assign individual as case

The logistic model

Define the penetrance of a series of genotypes, X as $P(\text{Case}|X)$, and the prevalence of a binary trait as $P(\text{Case})$.

A logistic function is used to calculate the penetrance for each X

$$P(\text{Case}|X) = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$$

Where X is comprised of genotypes at individual causal sites (coded for inheritance model).

β values are the log-odds ratios for each variant position. The α value is used to control the overall prevalence of the trait. $\alpha = \frac{f_0}{1-f_0}$ where f_0 is the $P(\text{Case}|X = 0)$

The odds ratio

In the logistic model we call the β values log-odds ratios. If we define the odds-ratio as:

$$OR_i = \frac{P(Y = 1|X_i = 1, \mathbf{Z})/P(Y = 0|X_i = 1, \mathbf{Z})}{P(Y = 1|X_i = 0, \mathbf{Z})/P(Y = 0|X_i = 0, \mathbf{Z})}$$

Then in the logistic model

$$exp^{\beta_i} = OR_i$$

Notice that the OR is calculated conditional on the other covariates in the model, $\mathbf{Z}$

Approach 1

Set the β values (or randomly draw them), and set a desired population prevalence, K.

Search over α to minimize $|\hat{K} - K|$, where $\hat{K}$ results from α

For each individual draw from $Unif(0, 1)$ and assign case status if draw $>$ P(Case)

Approach 2

Use the Population Attributable Risk (PAR) model.

Set f_0 and PAR_i for variant site i with frequency p_i, where $\sum PAR_i = PAR$. Then the Genotypic Relative Risk (GRR) is

$$GRR_i = \frac{PAR_i}{(1 - PAR_i)p_i} + 1$$

And

$$P(\text{Case}|X) = \frac{f_0 \prod_{i=1}^{m} GRR_i^X}{(1 - f_0) + f_0 \prod_{i=1}^{m} GRR_i^X}$$

Again, for each individual draw from $Unif(0, 1)$ and assign case status if draw $>$ P(Case)

A simple Example

Assume an additive model with 2 independent causal loci

$\beta_1 = 2, \quad \beta_2 = 0.25$

Set $P(\text{Case}|\text{No mutations}) = 0.1$

$$P(\text{Case}|X_1, X_2) = \frac{exp^{log(\frac{0.1}{0.9})+2X_1+0.25X_2}}{1 + exp^{log(\frac{0.1}{0.9})+2X_1+0.25X_2}}$$

Genotype	P(case)
(0, 0)	0.1
(1, 0)	0.45
(2, 0)	0.86
(0, 1)	0.12
(0, 2)	0.15
(1, 1)	0.51
(2, 1)	0.89
(1, 2)	0.58
(2, 2)	0.91

Software resources

SEQPower (http://bioinformatics.org/spower/simtraits) (Incorporates SimRare)

SeqSIMLA (http://seqsimla.sourceforge.net/)

phenosim (http://evoplant.uni-hohenheim.de/doku.php?id=software:software)

simuRareVariants (SRV) (http://simupop.sourceforge.net/Cookbook/SimuRareVariants)

Website with tons of Popgen simulation resources https://popmodels.cancercontrol.cancer.gov/gsr/packages/

Lab 3

- Use *ms* to simulate population growth, population structure, and recombination
- Evaluate how SFS and other sequence summaries change under each demographic scenario