Simulating GWAS Populations under Specific Disease Models

Daniel Ben-Isvy
CSCI 2820 – April 6, 2021

GWAS Simulation Pipeline Overview

- GWAS simulation pipelines generate synthetic GWAS data with similar characteristics as real GWAS data
- They allow researchers to evaluate the performance of GWAS test statistics under different models of population evolution and disease incidence
- 4 steps:
  - Simulate the evolution of a population forwards through time
  - Specify a disease model
  - Sample a case/control cohort from the population according to the disease model
  - Compute GWAS test statistics on the case/control cohort in an attempt to identify the disease locus

Forward Simulation Overview

- Forward simulations can be used to create populations with realistic patterns of genetic variation
- They can include a wide variety of population genetic processes (e.g. growth, bottlenecks, substructure, migration), but we focus here on a constant size population with no substructure

Forward Simulation Steps

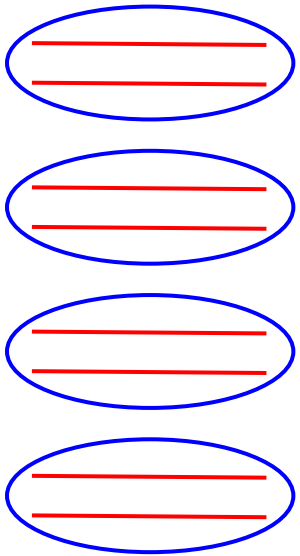
- Begin with a population of *N* genetically identical diploid individuals
- In each generation, create a population of *N* new individuals as follows:
  - Sample *N* pairs of individuals with replacement from the previous generation to reproduce
  - Determine the genetic material passed down to each offspring using **mathematical models of genetic events** such as mutation and recombination
- Run the simulation for as many generations as desired

Mathematical Models of Genetic Events

- The simplest mathematical models of genetic events specify a uniform rate at which those events occur across the genome
  - Example: uniform mutation rate of 1.1×10^{-8} mutations/base pair/individual/generation
  - Example: uniform recombination rate of 2.2×10^{-8} breakpoints/base pair/individual/generation
- More complex mathematical models of genetic events can specify “hotspots” of genetic activity in the genome
  - Example: recombination hotspots account for 1% of the total length of the genome but 60% of all recombination events
    - In this model, there are two recombination rates: one for inside hotspots and one for outside hotspots

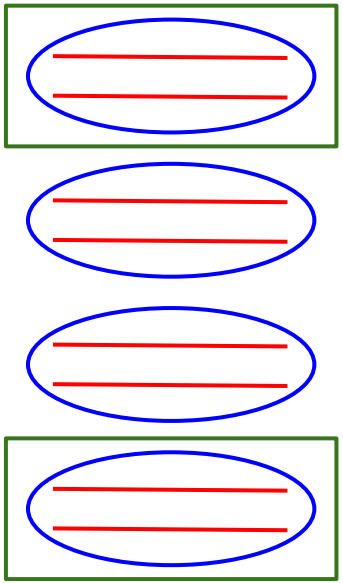
Forward Simulation: An Example (Initialization)

- Example with a population size of $N = 4$ and one chromosome of genetic material
- Begin with a population of N genetically identical diploid individuals



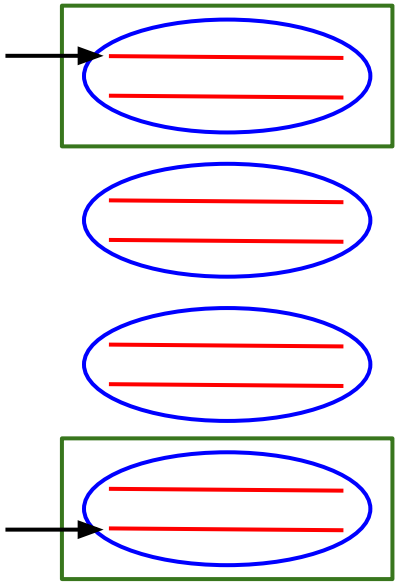
Forward Simulation: An Example (Generation 1)

- Randomly select the first pair of individuals to reproduce (random mating)



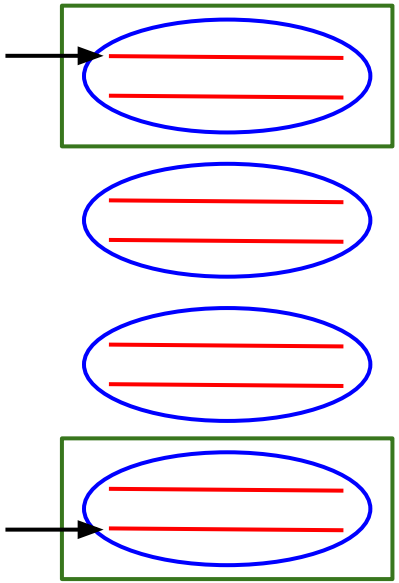
Forward Simulation: An Example (Generation 1)

- Randomly select a chromosome from each parent to be passed to the offspring (independent assortment)



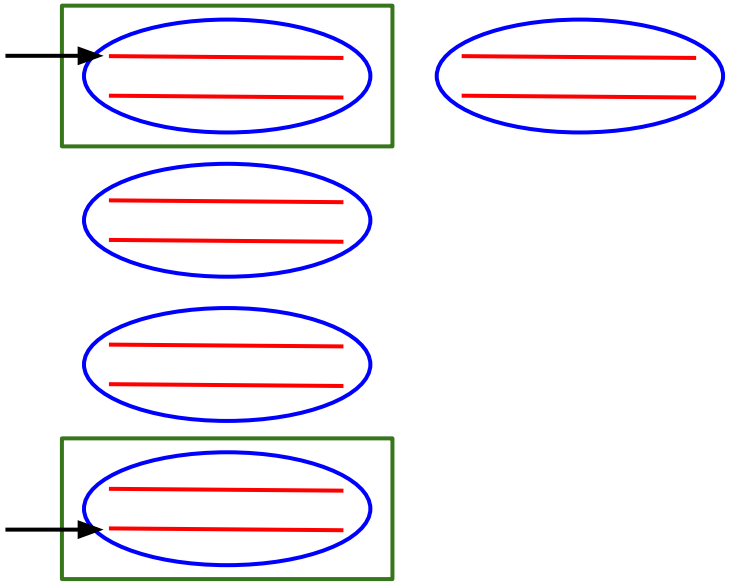
Forward Simulation: An Example (Generation 1)

- Randomly select recombination breakpoints according to the recombination model (none selected here)



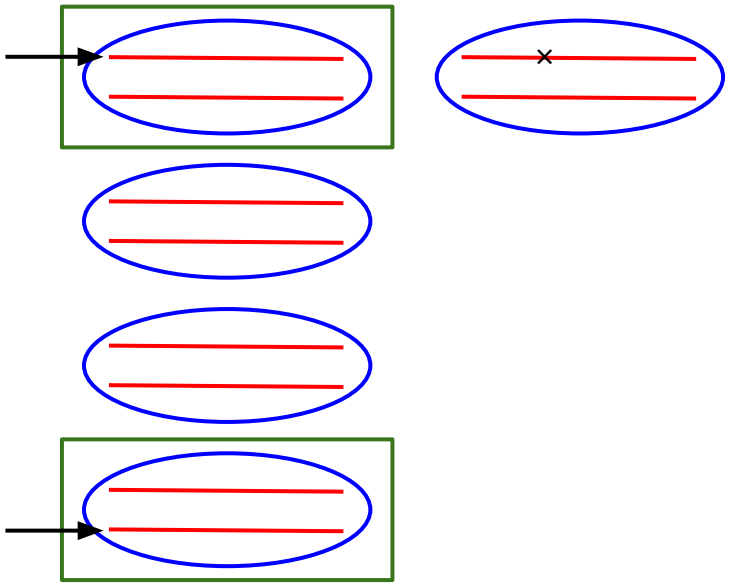
Forward Simulation: An Example (Generation 1)

- Pass the recombined genetic material to the offspring



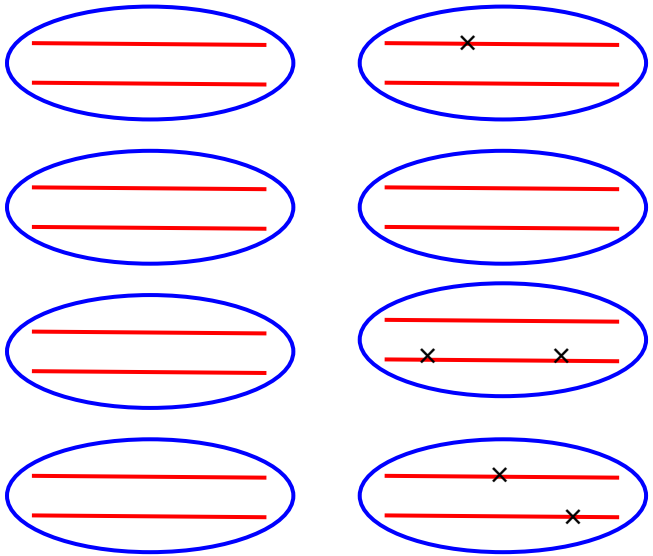
Forward Simulation: An Example (Generation 1)

- Randomly select mutation sites according to the mutation model



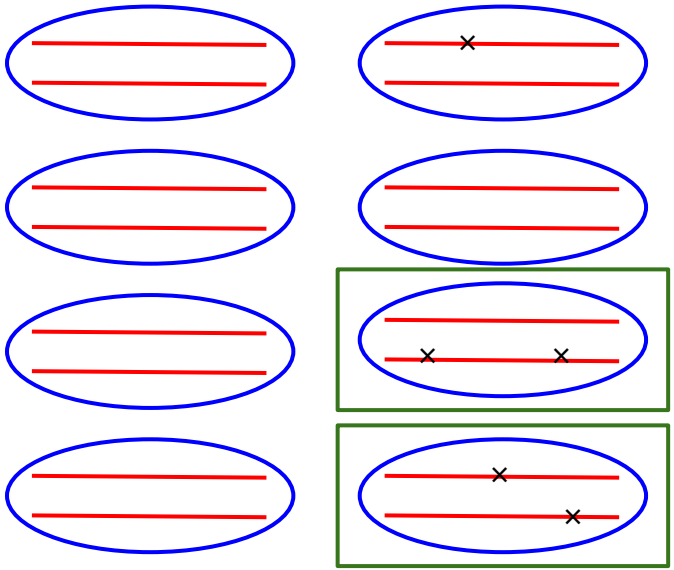
Forward Simulation: An Example (Generation 1)

- Repeat this process 3 more times to generate the rest of generation 1



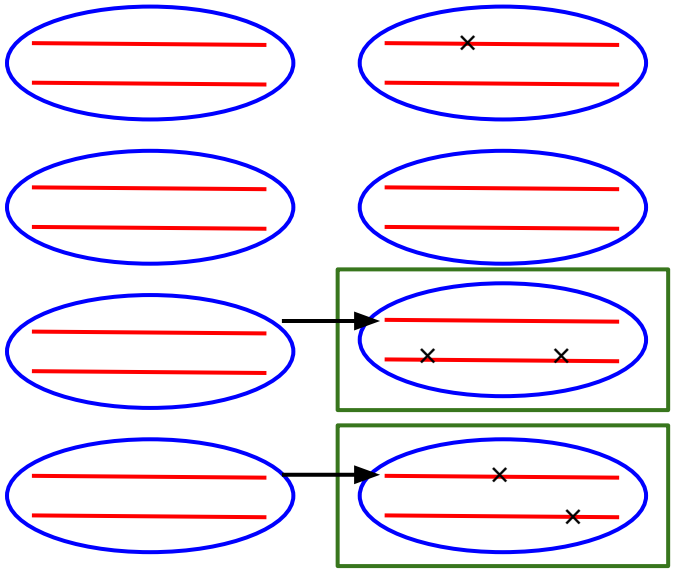
Forward Simulation: An Example (Generation 2)

- Randomly select the first pair of individuals to reproduce



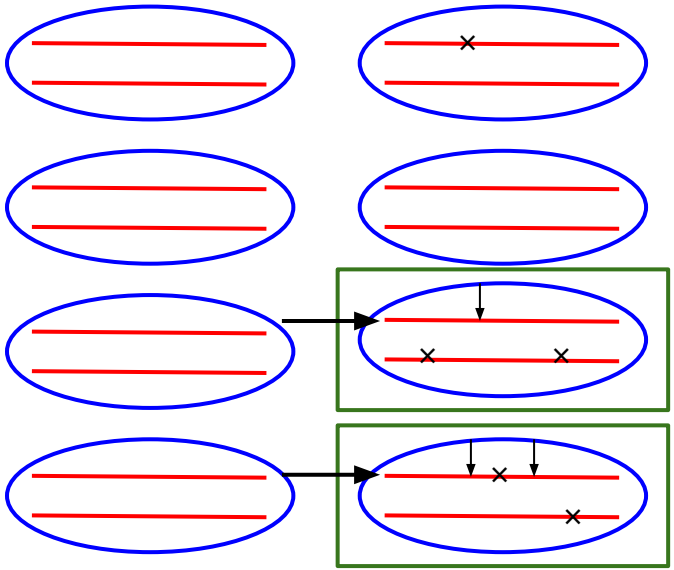
Forward Simulation: An Example (Generation 2)

- Randomly select a chromosome from each parent to be passed to the offspring



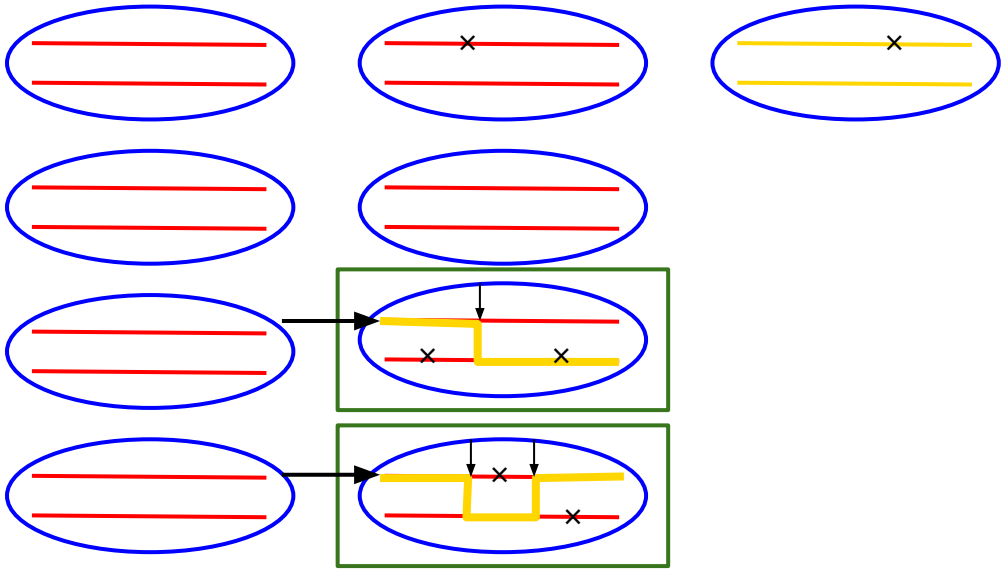
Forward Simulation: An Example (Generation 2)

- Randomly select recombination breakpoints according to the recombination model



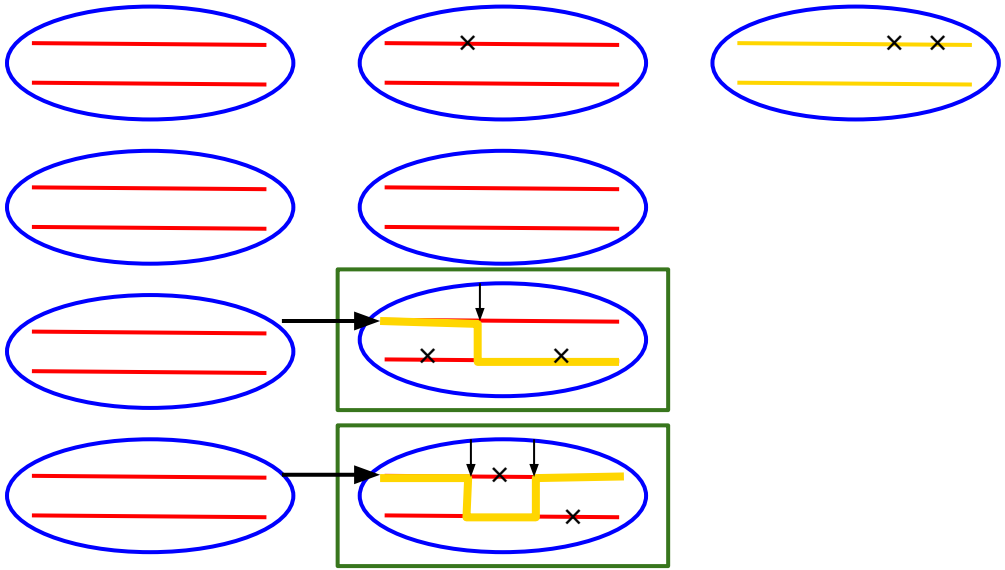
Forward Simulation: An Example (Generation 2)

- Pass the recombined genetic material to the offspring



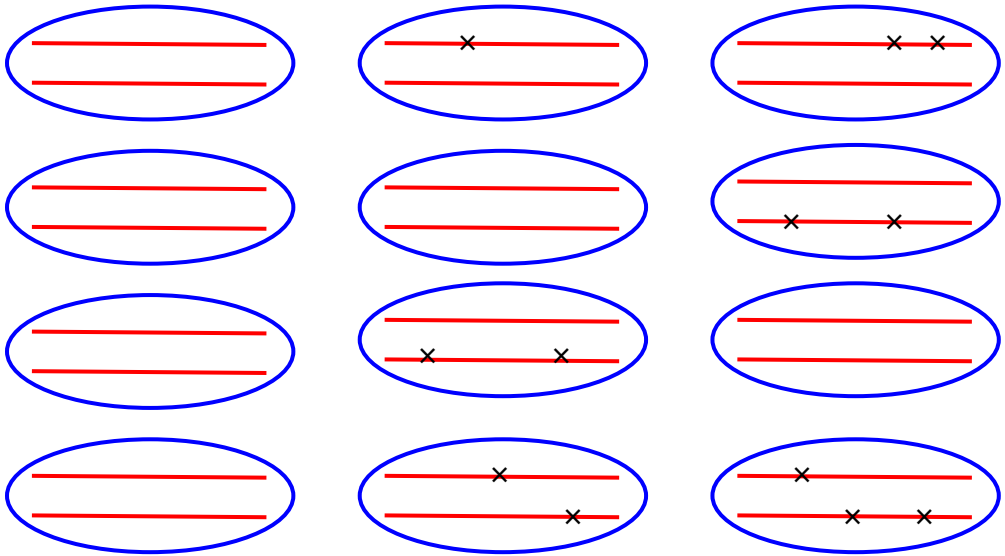
Forward Simulation: An Example (Generation 2)

- Randomly select mutation sites according to the mutation model



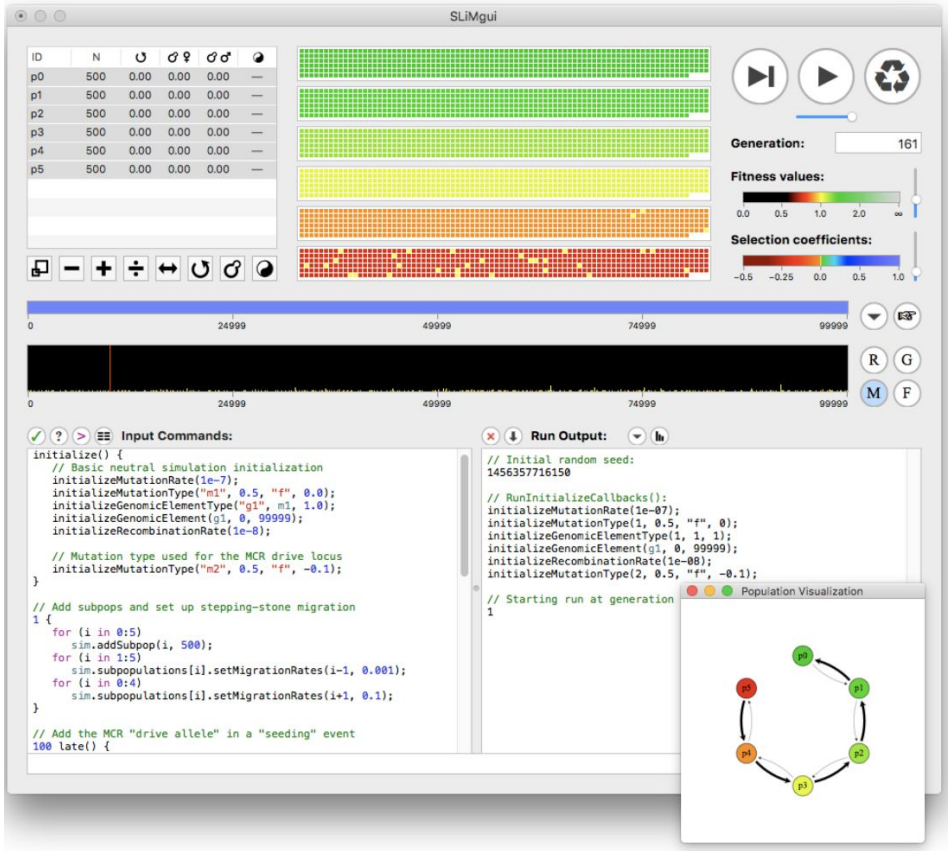
Forward Simulation: An Example (Generation 2)

- Repeat this process 3 more times to generate the rest of generation 2
- Run the simulation for as many generations as desired



Forward Simulation Software

- Many software packages exist to perform forward simulations efficiently, such as the evolutionary simulation framework **SLiM** (https://messerlab.org/slim)
- Designed specifically to perform evolutionary simulations
- Extensively documented
- Graphical user interface SLiMgui available (right)



Disease Model: Penetrance and GRR

- A = wild-type allele, a = disease allele
- **Penetrance**: the probability that an individual has the disease phenotype given that they have a particular genotype—denoted P(genotype)
- **Genotype relative risk**: how much more likely an individual with a particular genotype is to have the disease phenotype than an individual with the wild-type genotype—computed as a ratio of penetrances:
  - GRR(AA) = P(AA) / P(AA) = 1
  - GRR(Aa) = P(Aa) / P(AA)
  - GRR(aa) = P(aa) / P(AA)

Disease Model: Types of GRR Models

- **Additive disease model**: each copy of the disease allele increases disease risk by the same additive amount
  - $GRR(aa) = 2 \times GRR(Aa) - 1$
  - Most commonly used model for GWAS because complex disease phenotypes are often assumed to have approximately additive genetic determinants
- **Multiplicative disease model**: each copy of the disease allele increases disease risk by the same multiplicative amount
  - $GRR(aa) / GRR(Aa) = GRR(Aa) / GRR(AA) \rightarrow GRR(aa) = [GRR(Aa)]^2$
- **Dominant disease model**: all individuals with at least one copy of the disease allele have the same disease risk
  - $GRR(Aa) = GRR(aa)$
- **Recessive disease model**: individuals need two copies of the disease allele to have an increased risk of disease
  - $GRR(Aa) = 1$

Disease Model: Frequency and Prevalence

- **Disease allele frequency**: the frequency of the disease allele in the entire population (denoted q)
- **Disease prevalence**: the fraction of individuals in the entire population who have the disease phenotype (denoted p)

Disease Model: Parameterization

- A disease model can be fully specified by the 4 parameters GRR(Aa), GRR(aa), q, and p
- Assuming Hardy-Weinberg equilibrium, the penetrance for each genotype can be computed from these 4 parameters by solving the following linear system of 3 equations with 3 unknowns:
  - $p = [(1-q)^2 \times \mathbf{P(AA)}] + [2q(1-q) \times \mathbf{P(Aa)}] + [q^2 \times \mathbf{P(aa)}]$ — law of total probability
  - GRR(Aa) = **P(Aa)** / **P(AA)** → 0 = [GRR(Aa) × **P(AA)**] – **P(Aa)** — definition of GRR
  - GRR(aa) = **P(aa)** / **P(AA)** → 0 = [GRR(aa) × **P(AA)**] – **P(aa)** — definition of GRR
- If the population is not in Hardy-Weinberg equilibrium, the same method can still be used to compute the penetrances if the population genotype frequencies can be determined
  - They can no longer be computed from just the allele frequency

Disease Model: An Example Penetrance Calculation

- Let GRR(Aa) = 2, GRR(aa) = 3, p = 0.01, and q = 0.1
- System of equations:
  - $0.01 = [0.81 \times \mathbf{P(AA)}] + [0.18 \times \mathbf{P(Aa)}] + [0.01 \times \mathbf{P(aa)}]$
  - $0 = [2 \times \mathbf{P(AA)}] - \mathbf{P(Aa)} \rightarrow \mathbf{P(Aa)} = 2 \times \mathbf{P(AA)}$
  - $0 = [3 \times \mathbf{P(AA)}] - \mathbf{P(aa)} \rightarrow \mathbf{P(aa)} = 3 \times \mathbf{P(AA)}$
- Substitution:
  - $0.01 = [0.81 \times \mathbf{P(AA)}] + [0.18 \times [2 \times \mathbf{P(AA)}]] + [0.01 \times [3 \times \mathbf{P(AA)}]] \rightarrow 0.01 = 1.2 \times \mathbf{P(AA)} \rightarrow \mathbf{P(AA)} = 1/120$
  - $\mathbf{P(Aa)} = 2 \times (1/120) = 1/60$
  - $\mathbf{P(aa)} = 3 \times (1/120) = 1/40$

Sampling a Case/Control Cohort

- Randomly select a disease locus in the population from all loci with minor allele frequency between $q-\varepsilon$ and $q+\varepsilon$ for some small number ε
- Determine the number of case and control individuals to be sampled
- While the case and control groups are not both full:
  - Randomly select two haplotypes from the population, forming an individual
  - Determine the genotype G of this individual at the disease locus
  - The penetrance $P(G)$ gives the probability that this individual has the disease phenotype given their genotype at the disease locus
  - Based on this probability, randomly draw the phenotype of the individual
  - Add the individual to the appropriate group (case or control) if that group is not yet full
- Return the disease locus and the case/control cohort

Sampling a Case/Control Cohort: An Example

Controls (0/10)

Cases (0/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Based on the penetrance P(G), randomly draw the phenotype

Add to the appropriate study group if it is not full

Sampling a Case/Control Cohort: An Example

Controls (1/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Cases (0/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

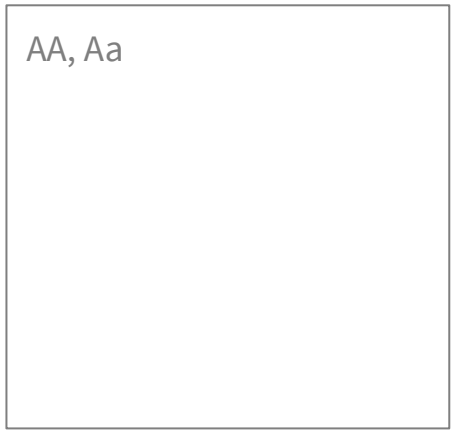
Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example



Controls (2/10)

Cases (0/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa

Controls (2/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa

Cases (1/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa

Controls (2/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa

Cases (2/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa

Controls (2/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa

Cases (3/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa

Controls (3/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa

Cases (3/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA

Controls (4/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa

Cases (3/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA

Controls (5/10)

Aa, aa, aa

Cases (3/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA

Controls (6/10)

Aa, aa, aa

Cases (3/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA

Controls (6/10)

Aa, aa, aa, AA

Cases (4/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa

Controls (7/10)

Aa, aa, aa, AA

Cases (4/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa

Controls (8/10)

Aa, aa, aa, AA

Cases (4/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa

Controls (8/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa, AA, aa

Cases (5/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa

Controls (9/10)

Aa, aa, aa, AA, aa

Cases (5/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa

Controls (9/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa, AA, aa, Aa

Cases (6/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa, AA, aa, Aa

Cases (6/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa, AA, aa, Aa

Cases (6/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Not added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa

Cases (6/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Not added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Aa, aa, aa, AA, aa, Aa, aa

Cases (7/10)

To sample:
10 controls, 10 cases

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa

Cases (8/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa

Cases (8/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Not added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa

Cases (8/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

AA

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Not added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa

Cases (8/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Control

Add to the appropriate study group if it is not full

Not added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa, Aa

Cases (9/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

Aa

Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example

AA, Aa, aa, AA, AA, AA, Aa, Aa, aa, Aa

Controls (10/10)

Aa, aa, aa, AA, aa, Aa, aa, Aa, Aa, aa

Cases (10/10)

To sample:
10 controls, 10 cases

P(AA)	1/8
P(Aa)	1/4
P(aa)	1/2

Penetrances

Steps:

Randomly select two haplotypes from the population and determine the genotype G at the disease locus

aa

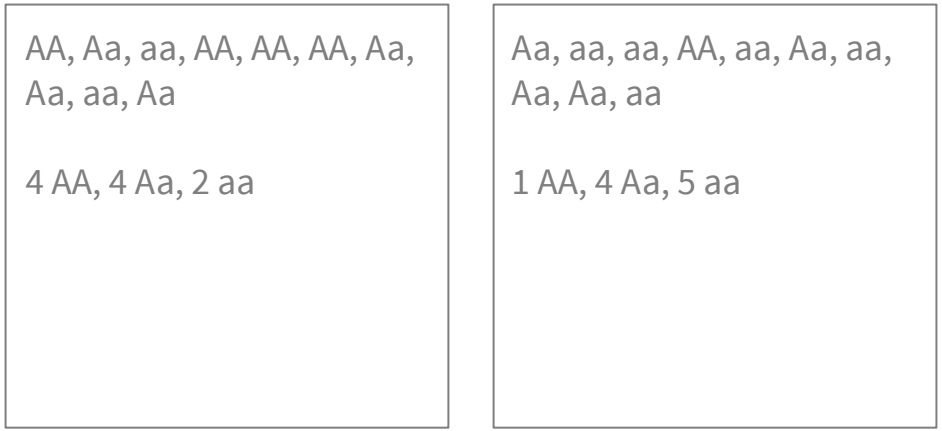
Based on the penetrance P(G), randomly draw the phenotype

Case

Add to the appropriate study group if it is not full

Added

Sampling a Case/Control Cohort: An Example



Controls (10/10) Cases (10/10)

Final case/control cohort

GWAS Test Statistics: The χ^2 Test

- Determines whether there is a statistically significant difference between the observed and expected frequencies in the contingency table
- Null hypothesis: no association between the disease phenotype and the genotype at the locus of interest

- Test statistic: $\chi^2 = \sum_{cells} \frac{(observed - expected)^2}{expected}$

- For a 2 × 3 contingency table, the test statistic is χ^2-distributed with 2 degrees of freedom

GWAS Test Statistics: The χ^2 Test

- Contingency table of observations:

Genotype	AA	Aa	aa	Total
Controls	*a*	*b*	*c*	*a+b+c*
Cases	*d*	*e*	*f*	*d+e+f*
Total	*a+d*	*b+e*	*c+f*	*a+b+c+d+e+f*

- Under the null hypothesis, we expect the genotype and phenotype to be independent
  - Example: (expected # AA controls) = (fraction of individuals with AA genotype) × (number of controls) = [(a+d) / (a+b+c+d+e+f)] × (a+b+c)

GWAS Simulation Pipeline Review

- 4 steps in a GWAS simulation pipeline:
  - Simulate the evolution of a population forwards through time
  - Specify a disease model
  - Sample a case/control cohort from the population according to the disease model
  - Compute GWAS test statistics on the case/control cohort in an attempt to identify the disease locus