EEB 70901 Population Genetics

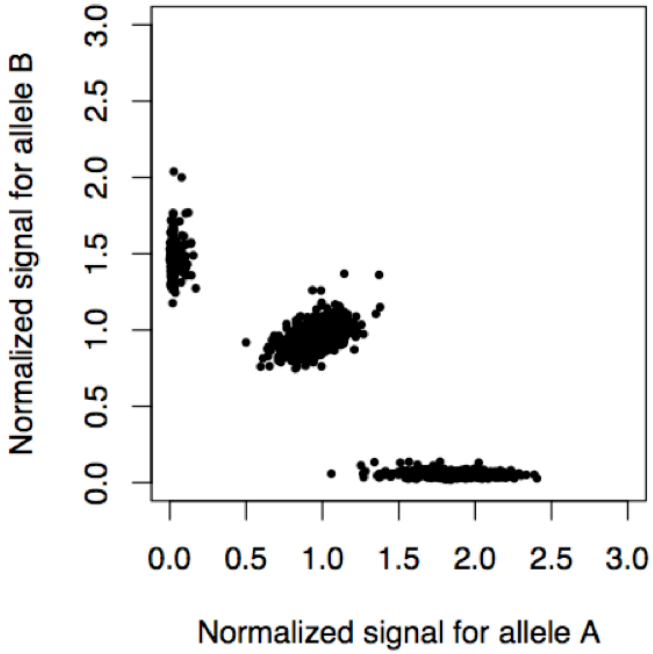
Lecture 2:
Genetic Variation

Professor Mike Hickerson
Department of Biology - City College
CUNY GC subprogram in EEB

SNP genotyping arrays (a.k.a. “SNP chips”)

Single Nucleotide Polymorphism (SNP): A locus where a single nucleotide site has more than one allele in the population

Probes are built to interrogate the genotype at each locus



Current arrays can genotype >1 million SNPs in the genome

Consider a simple data set

Example: 3 SNPs from 6 individuals

```
Individual ID| Genotypes
42724 C G / T T / T T / ...
42727 G G / C T / T T / ...
44163 C G / T T / T T / ...
...
45501 G G / T T / T T / ...
48693 G G / T T / G T / ...
14215 G G / C T / G T / ...
```

Consider a simple data set

Example: 3 SNPs from 6 individuals

```
Individual ID| Genotypes
42724 C G / T T / T T / ...
42727 G G / C T / T T / ...
44163 C G / T T / T T / ...
...
45501 G G / T T / T T / ...
48693 G G / T T / G T / ...
14215 G G / C T / G T / ...
```

We can summarize the data by counting each genotype:

Genotype counts

SNP name	Allele A	Allele a	n_{AA}	n_{Aa}	n_{aa}	n
SNP_A-4213906	C	G	16	85	113	214
SNP_A-2157996	C	T	25	93	104	222
SNP_A-4228006	G	T	3	48	165	216

NOTE: Not all genotype counts sum to 222 because some individual genotypes are not called ("missing")

Summarizing using frequencies

Genotype counts

SNP name	Allele A	Allele a	n_{AA}	n_{Aa}	n_{aa}	n
SNP_A-4213906	C	G	16	85	113	214
SNP_A-2157996	C	T	25	93	104	222
SNP_A-4228006	G	T	3	48	165	216

Genotype frequencies

SNP name	$X = n_{AA}/n$	$Y = n_{Aa}/n$	$Z = n_{aa}/n$
SNP_A-4213906	0.07	0.40	0.53
SNP_A-2157996	0.11	0.42	0.47
SNP_A-4228006	0.014	0.22	0.76

converting genotype frequencies to allele frequencies

Allele frequencies

- Let p = frequency of chromosomes with allele *A*
- Let q = frequency of chromosomes with allele *a*

Genotype frequencies

SNP name	$X = n_{AA}/n$	$Y = n_{Aa}/n$	$Z = n_{aa}/n$
SNP_A-4213906	0.07	0.40	0.53
SNP_A-2157996	0.11	0.42	0.47
SNP_A-4228006	0.014	0.22	0.76

converting genotype frequencies to allele frequencies

Allele frequencies

- Let p = frequency of chromosomes with allele *A*
- Let q = frequency of chromosomes with allele *a*

Genotype frequencies

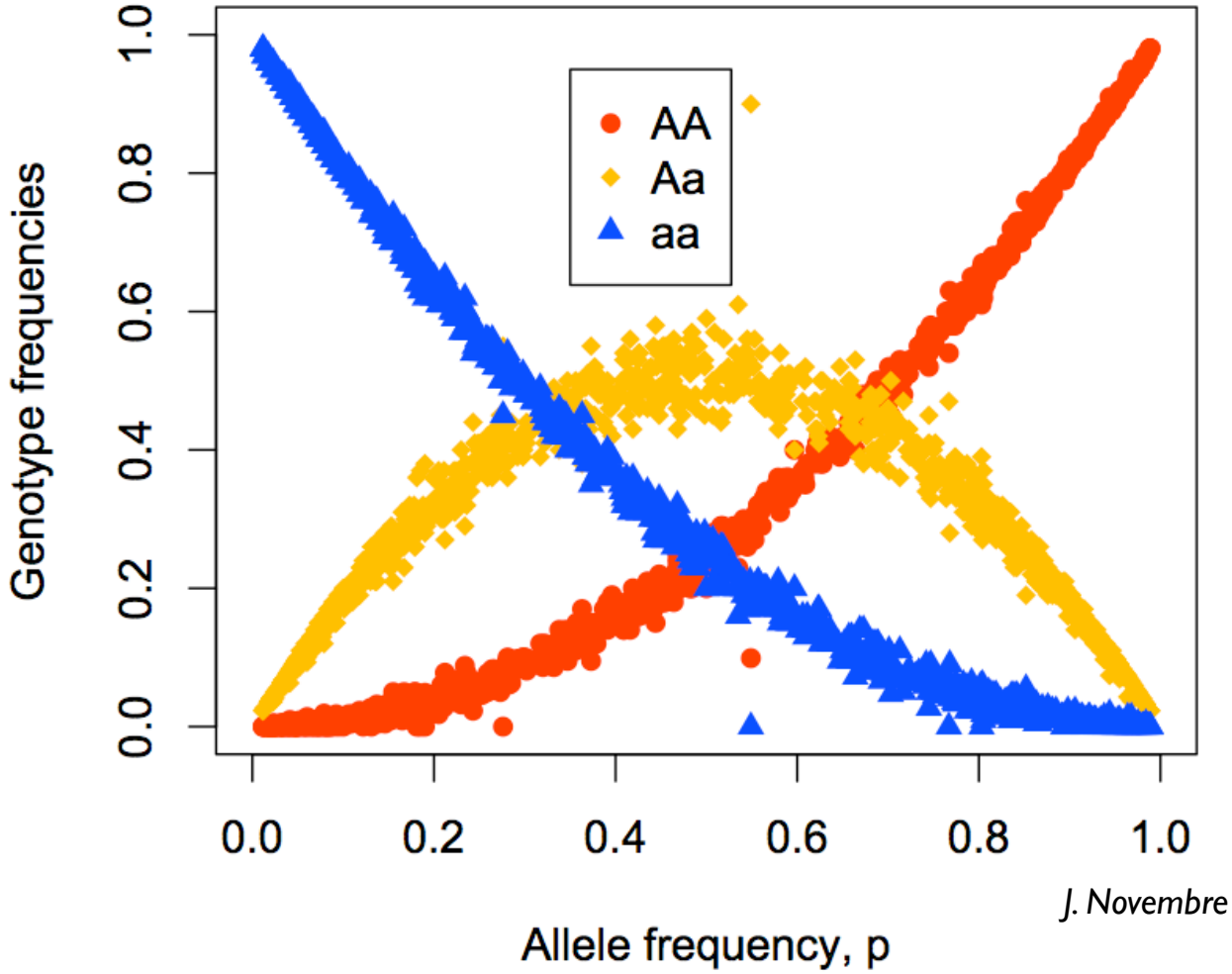
SNP name	$X = n_{AA}/n$	$Y = n_{Aa}/n$	$Z = n_{aa}/n$
SNP_A-4213906	0.07	0.40	0.53
SNP_A-2157996	0.11	0.42	0.47
SNP_A-4228006	0.014	0.22	0.76

Allele frequencies

SNP name	$p = (2n_{AA} + n_{Aa})/2n$	$q = (2n_{aa} + n_{Aa})/2n$
SNP_A-4213906	0.27	0.73
SNP_A-2157996	0.32	0.68
SNP_A-4228006	0.124	0.876

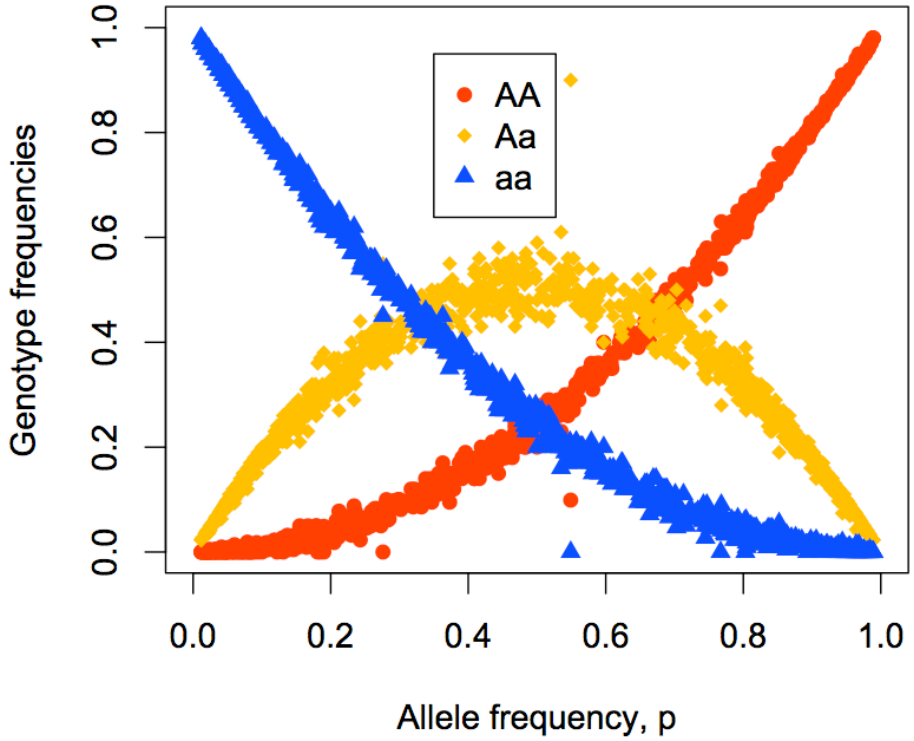
Now... let's see if there is any connection between genotype frequencies and allele frequencies.

For each locus, we plot the frequency of AA, Aa, and aa, as a function of p. (i.e. 3 points vertically aligned, with 1 of each color, for each SNP).



Now... let's see if there is any connection between genotype frequencies and allele frequencies.

For each locus, we plot the frequency of AA, Aa, and aa, as a function of p. (i.e. 3 points vertically aligned, with 1 of each color, for each SNP).



Why do we see these striking regularities? Is there some underlying law of nature governing this? Can we build a theory that explains these observations?

**All models are Wrong,
but some models are useful
(or more useful than others,
or less wrong than others)
1976, 1978 & 1987**

George Box
RIP 2013

Model –approximation of reality

identifying parameters

estimating parameters

testing (goodness of fit; data-model)

Verbal Models

Toy Models-does the verbal model hold up? (in principle)

If yes, then we can Identify key parameters (e.g. population size, mutation rate, migration rate)

Complex mathematical/statistical Models

Estimate the parameters values with genomic data

Verbal Models

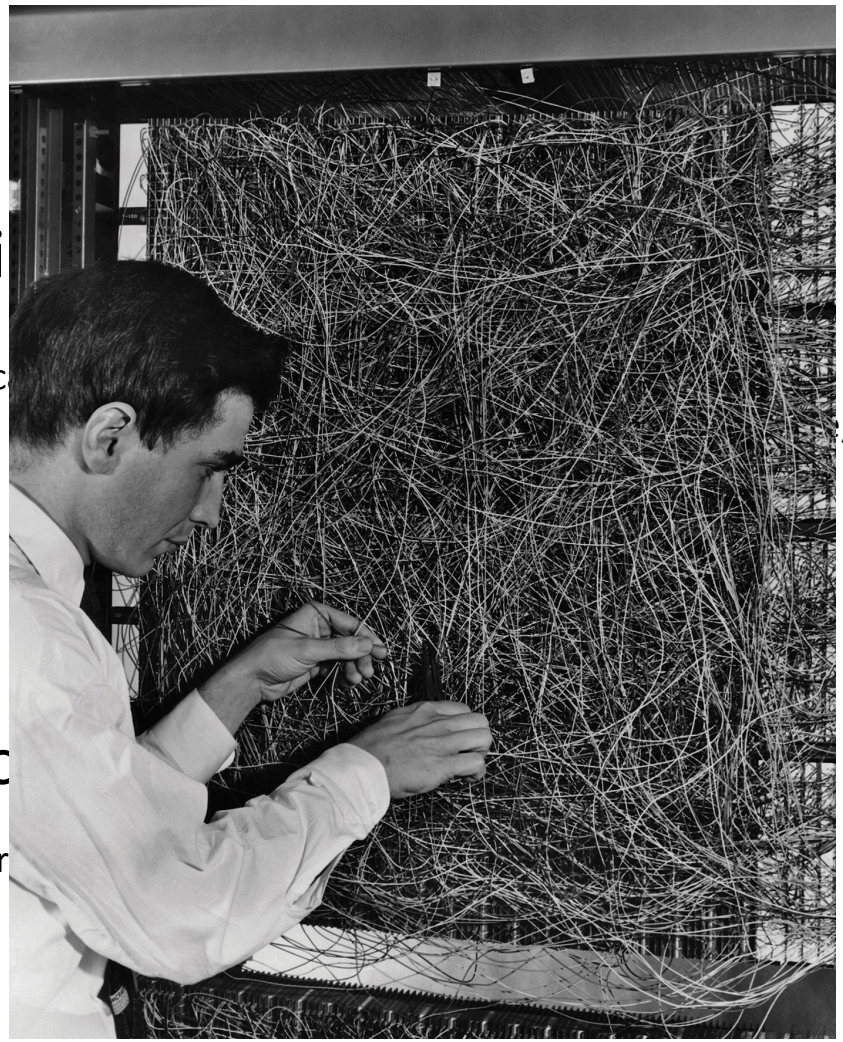
Toy Models-does the
(in princi

If yes, then we c

Complex mathematic

Estimate the par

Machine learning



Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Let’s make a simple model:

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Let's make a simple model:

Begin with genotype frequencies X,Y,Z.

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Let's make a simple model:

Begin with genotype frequencies X,Y,Z (AA, Aa, and aa).

Then we'll compute the frequencies of gametes we expect to be generated during reproduction.

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Let's make a simple model:

Begin with genotype frequencies X,Y,Z.

Then we'll compute the frequencies of gametes we expect to be generated during reproduction.

Then we'll assume the gametes are mated at random, and compute the expected frequencies of genotypes in the next generation.

Building theoretical model for expected genotype frequencies

What does it take to create the AA, Aa, and aa individuals?

Gametes!

Let's make a simple model:

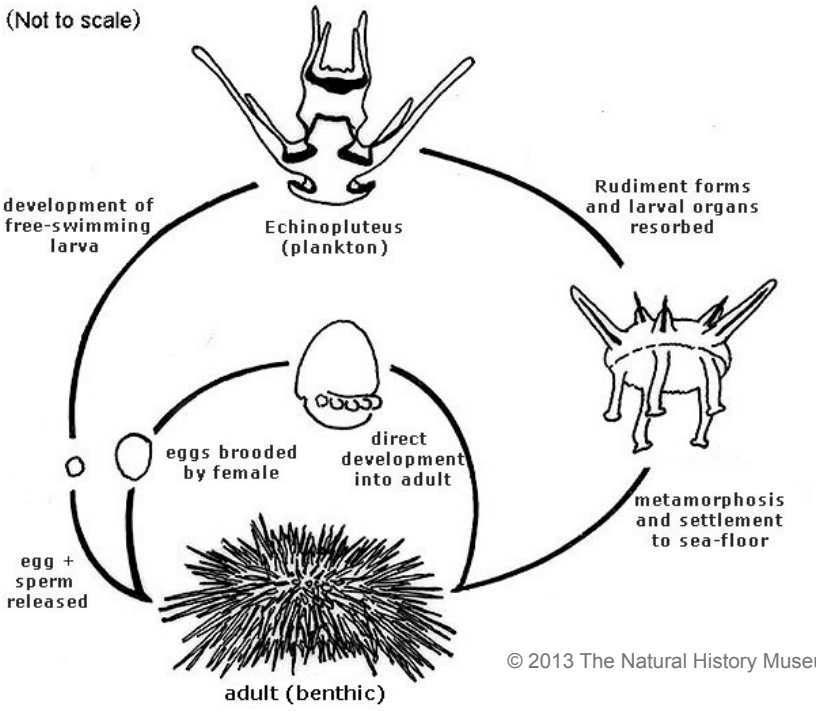
Begin with genotype frequencies X,Y,Z.

Then we'll compute the frequencies of gametes we expect to be generated during reproduction.

Then we'll assume the gametes are mated at random, and compute the expected frequencies of genotypes in the next generation.

Question

Is this simple theory going to be enough to explain the patterns in our data?



Building theoretical model for expected genotype frequencies - part 2

First, let's remind ourselves about Mendel:

From Mendel's law of segregation of alleles...

Genotype	*A* gametes	*a* gametes
AA	100%	0
Aa	50%	50%
aa	0	100%

Thus, if a population has genotype frequencies X, Y, Z, we have:

Expected gamete frequencies

Gamete	Expected frequency
A	$1.0X + 0.5Y = p$
a	$1.0Z + 0.5Y = q$

Building theoretical model for expected genotype frequencies - part 3

Expected gamete frequencies

Gamete	Expected frequency
A	$1.0X + 0.5Y = p$
a	$1.0Z + 0.5Y = q$

And if gametes are joining each other at random:

Expected genotype frequencies in next generation

Gamete 1	Gamete 2	Genotype	Expected frequency
A	A	AA	$p \times p = p^2$
A	a	Aa	$p \times q = pq$
a	A	Aa	$q \times p = qp$
a	a	aa	$q \times q = q^2$

Building theoretical model for expected genotype frequencies - part 4

And so in sum, the expected genotype frequencies in the next generation are:

Expected genotype frequencies in next generation

Genotype	Expected frequency
AA	$X' = p \times p = p^2$
Aa	$Y' = p \times q + q \times p = 2pq$
aa	$Z' = q \times q = q^2$

Building theoretical model for expected genotype frequencies - part 4

And so in sum, the expected genotype frequencies in the next generation are:

Expected genotype frequencies in next generation

Genotype	Expected frequency
AA	$X' = p \times p = p^2$
Aa	$Y' = p \times q + q \times p = 2pq$
aa	$Z' = q \times q = q^2$

But wait, at this point ***p*** and ***q*** are the allele frequencies that existed in the current generation and we've calculated expected genotype frequencies for the next generation.

To understand the graph, we still need to connect how genotype and allele frequencies compare to each other in the same generation.

So let's calculate the allele frequency in the next generation... p' and q':

Allele frequency in the next generation

$$p' = X' + 0.5Y'$$
$$q' = Z' + 0.5Y'$$

So let's calculate the allele frequency in the next generation... p' and q':

Allele frequency in the next generation

$$p' = X' + 0.5Y' = p^2 + 0.5(2pq) = p$$
$$q' = Z' + 0.5Y' = q^2 + 0.5(2pq) = q$$

So let's calculate the allele frequency in the next generation... p' and q':

Allele frequency in the next generation

$$p' = X' + 0.5Y' = p^2 + 0.5(2pq) = p$$
$$q' = Z' + 0.5Y' = q^2 + 0.5(2pq) = q$$

Two implications of our theory:

- We have expected proportions $X' = (p')^2$, $Y' = 2p'q'$ and $Z' = (q')^2$
- Allele frequencies don't change! $p' = p$ and $q' = q$. If our theory is true, the population would be at *equilibrium* with regards to allele frequency.

So let's calculate the allele frequency in the next generation... p' and q':

Allele frequency in the next generation

$$p' = X' + 0.5Y' = p^2 + 0.5(2pq) = p$$
$$q' = Z' + 0.5Y' = q^2 + 0.5(2pq) = q$$

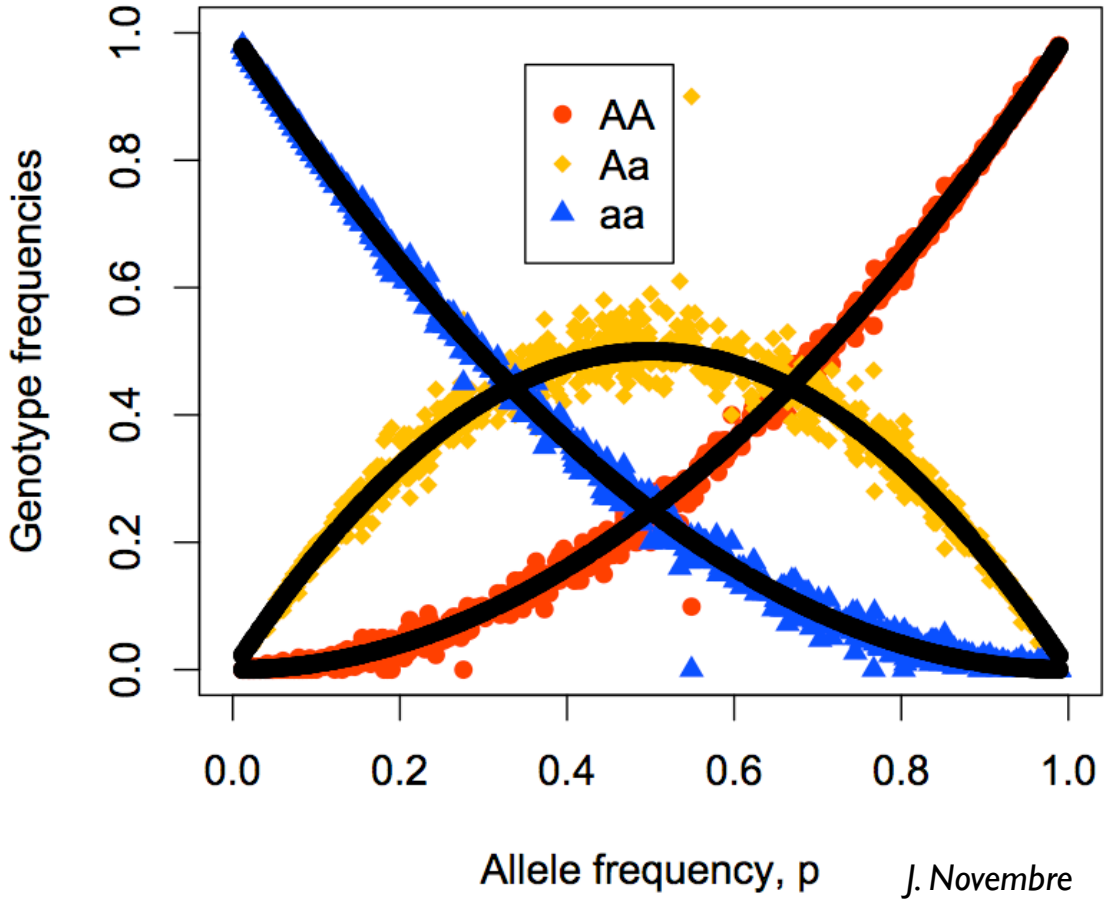
Two implications of our theory:

- We have expected proportions $X' = (p')^2$, $Y' = 2p'q'$ and $Z' = (q')^2$
- Allele frequencies don't change! $p' = p$ and $q' = q$. If our theory is true, the population would be at *equilibrium* with regards to allele frequency.

Question...

Do our theoretical calculations check out?

Theory vs Data



Black lines are plots of p^2, $2pq$, and q^2.

- The basic proportions p^2, $2pq$, and q^2 are called the Hardy-Weinberg proportions.

- The fact that they hold in many data-sets is sometimes called the **Hardy-Weinberg** Law

- The fact that the allele frequency does not change is why the results are often called Hardy-Weinberg equilibrium (HWE).

Note that our derivation assumed random union of gametes, which is an outcome of random mating.

We also implicitly assumed:

- Diploid individuals with sexual reproduction
- Only 2 alleles at the locus
- Generations are discrete and non-overlapping
- Males and females are equivalent
- Assumed the genotypes are measured accurately

Note that our derivation assumed random union of gametes, which is an outcome of random mating.

We also implicitly assumed:

- Diploid individuals with sexual reproduction
- Only 2 alleles at the locus
- Generations are discrete and non-overlapping
- Males and females are equivalent
- Assumed the genotypes are measured accurately

We also ignored any potential effects of:

- Natural selection
- Migration of individuals into the population
- Mutation
- Genetic drift (equivalent to assuming population size is very large)

Yet, as we saw from the comparison to real human SNP data, by eye, it seems to work quite well for most loci!

Complications from dominance

What if you cannot observe all three genotypes clearly?

Definition

Complete Dominance = the heterozygote is indistinguishable from one of the homozygotes

Complete dominance

Genotype	Observed phenotype
AA	1
Aa	1
aa	0

In this case we say the 1 phenotype is dominant and the 0 phenotype is *recessive*. We also say A is dominant over a and we say a is recessive to A. If we let R be the frequency of the 0 phenotype, we can still estimate q by *assuming HWE* and taking $\hat{q} = \sqrt{R}$ and $\hat{p} = 1 - \hat{q}$.

Reasons for lack of HW equilibrium

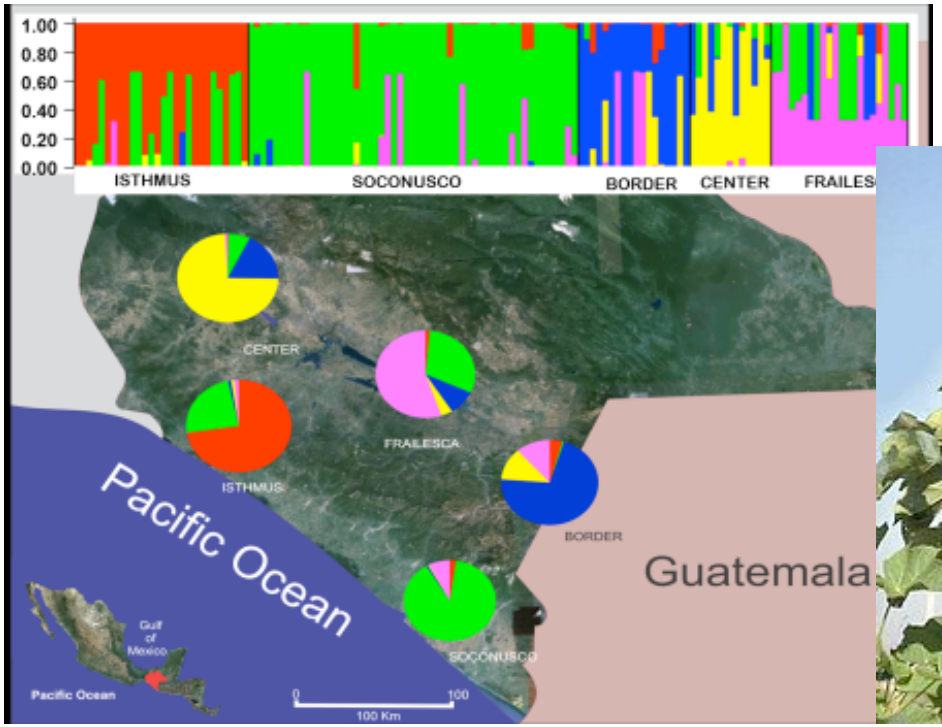
Assortative or dis-assortative mating (non-random mating)

- Assortative - mating more likely with similar genotypes (causing excess homozygotes). Can result from **inbreeding**
- dis-assortative - mating less likely with similar genotypes (causing excess heterozygotes)

Reasons for lack of HW equilibrium

Population Structure

causes non-random mating. e.g. If the allele frequencies are diverged, the total heterozygosity will be lower than HWE predictions



Ovando-Medina 2011

Reasons for lack of HW equilibrium

Selection

selection against heterozygotes, certain alleles, homozygotes can all cause genotype frequencies to deviate from HWE expectations

Conclusions

The importance of Hardy-Weinberg proportions:

- A general rule explaining how genotype frequencies and allele frequencies are related.

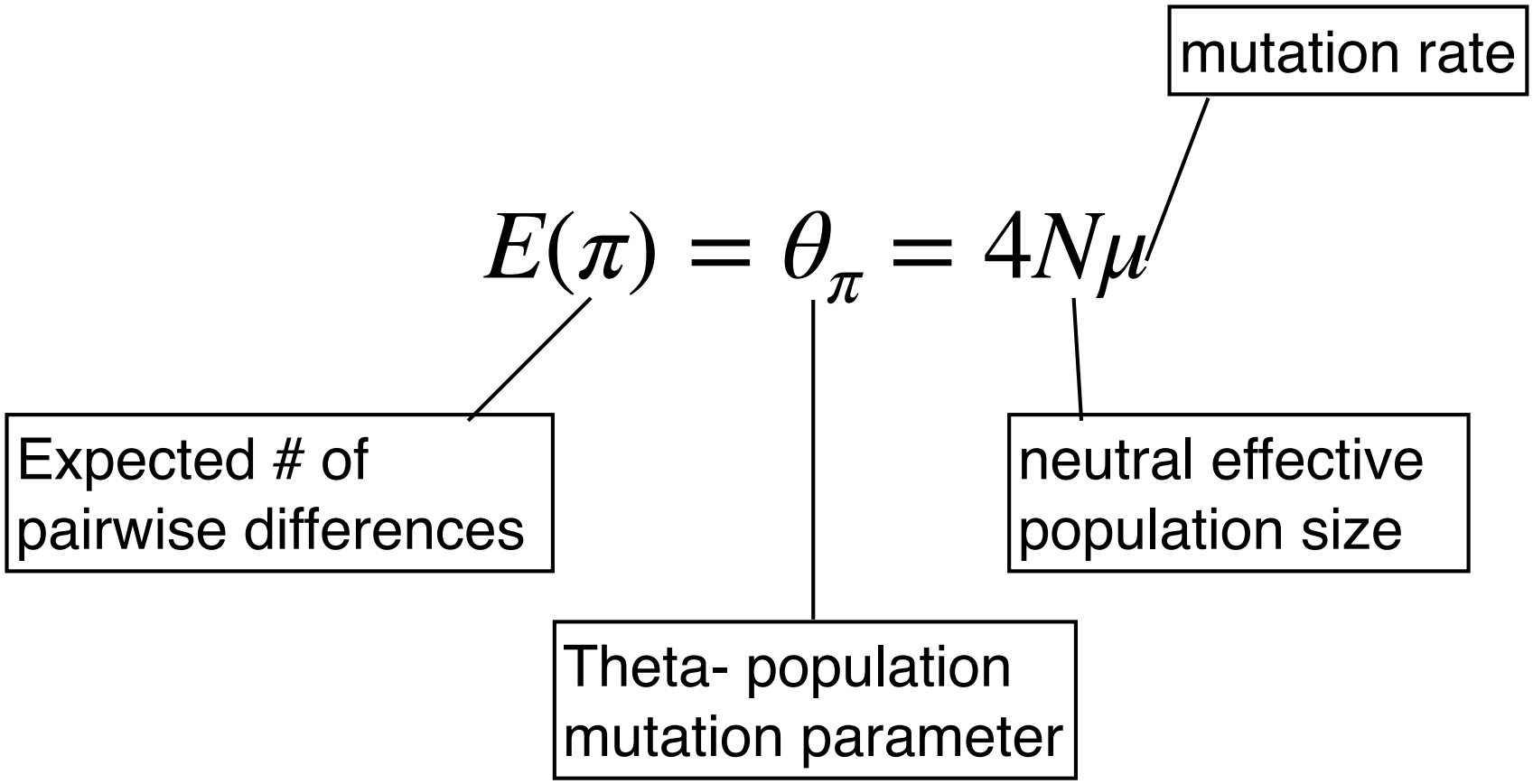
Conclusions

The importance of Hardy-Weinberg proportions:

- A general rule explaining how genotype frequencies and allele frequencies are related.
- No matter what X, Y, and Z are, with one generation of random mating, the genotype frequencies will fall into the HW proportions - this partially explains its pervasiveness

Measures of Genetic Variability

"Summary statistic" — a calculation that summarizes the data
— often corresponds to a parameter we want to estimate



ADH locus

	D. melanogaster												*D. simulans*						*D. yakuba*												
pos.	con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	a	b	c	d	e	f	g	h	i	j	k	l	NS
781	G	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NS
789	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
808	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
816	G	T	T	T	T	-	-	-	-	-	-	-	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
834	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	S
859	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	NS
867	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	A	G	G	G	G	G	G	S
870	C	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
950	G	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
974	G	-	-	-	-	-	-	-	-	-	-	-	-	T	-	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	S
983	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	S
1034	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		C	C	C	C	C	-	-	C	-	C	C	S
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	S
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	S
1089	C	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	NS
1101	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	NS
1127	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	S
1160	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	S

π for D. simulans

$$\pi = (1/15) * ((2+1+1+1+0) + (3+3+3+2) + (0+0+1) + (0+1) + (1)) = 1.26$$

usually normalized for # of base-pairs

π = 1.26/397 = 0.0032

Measures of Genetic Variability

Number of segregating sites S = total number of sites that are polymorphic

$$\theta_W$$

slowly grows with sample size n. We can scale this by # individuals n,

$$\widehat{\theta}_W = \frac{S}{\sum_{i=n-1}^{1} 1/i}$$

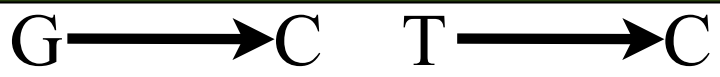
Site frequency Spectrum

a common way to summarize allele frequencies

Example: 3 SNPs from 6 Swiss individuals:

```
Individual ID| Genotypes
42724 C G / T T / T T / ...
42727 G G / C T / T T / ...
44163 C G / T T / T T / ...
...
45501 G G / T T / T T / ...
48693 G G / T T / G T / ...
14215 G G / C T / T T / ...
```

Site frequency Spectrum

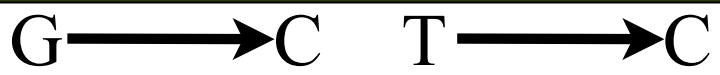


The proportion of SNPs w/ 2 derived alleles of frequency is 0.67

Example: 3 SNPs from 6 Swiss individuals:

```
Individual ID| Genotypes
42724 C G / T T / T T / ...
42727 G G / C T / T G / ...
44163 C G / T T / G T / ...
...
45501 G G / T T / T T / ...
48693 G G / T T / G T / ...
14215 G G / C T / G T / ...
```

Site frequency Spectrum



The proportion of SNPs w/ 2 derived alleles of frequency is 0.67

Example: 3 SNPs from 6 Swiss individuals:

```
Individual ID| Genotypes
42724 C G / T T / T T / ...
42727 G G / C T / T G / ...
44163 C G / T T / G T / ...
...
45501 G G / T T / T T / ...
48693 G G / T T / G T / ...
14215 G G / C T / G T / ...
```

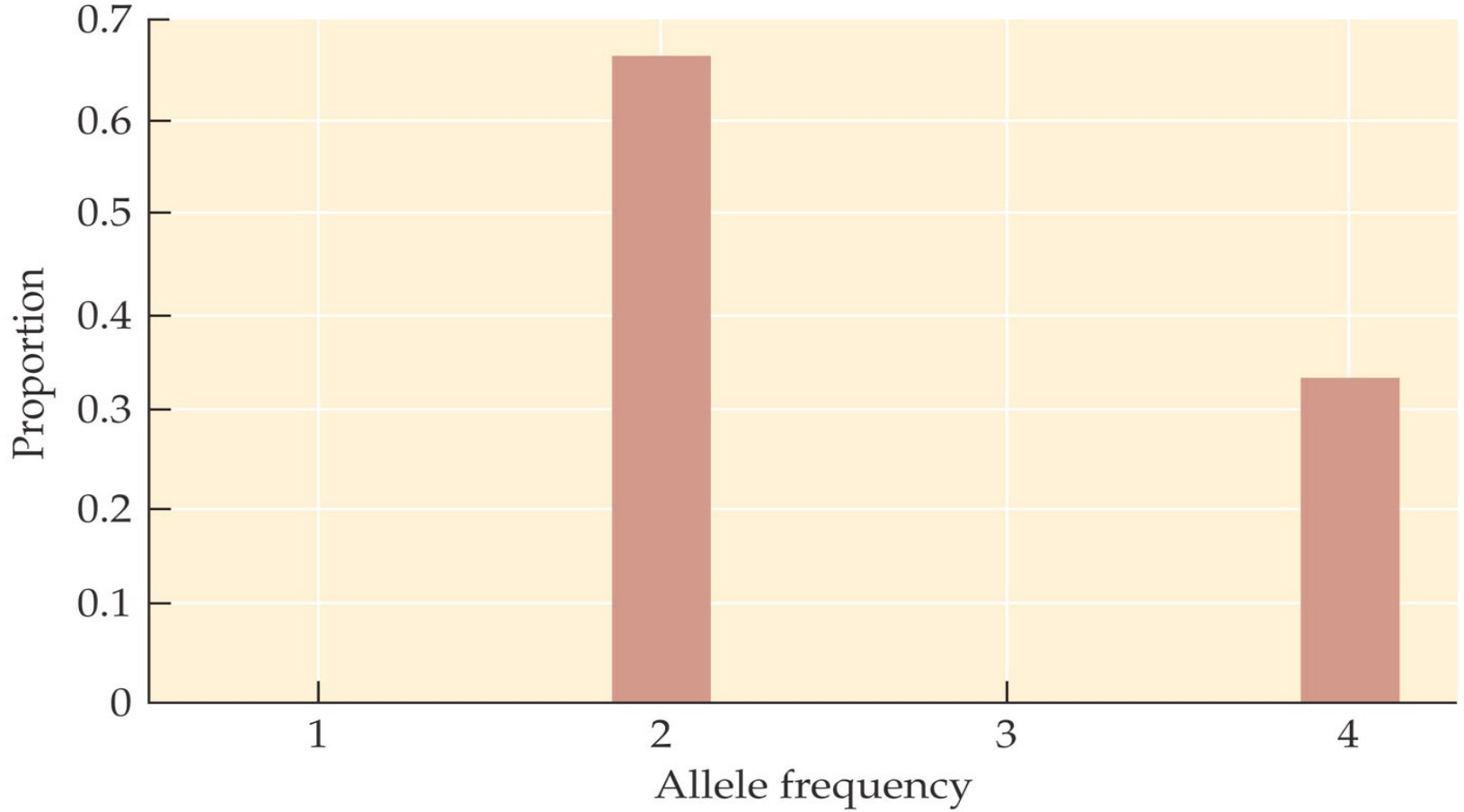
(or you could say % of SNPs w/ allele frequency of 0.33 is 0.67)

The proportion of SNPs w/ 4 derived alleles of frequency is 0.33

(or you could say % of SNPs w/ allele frequency of 0.67 is 0.33)

Site frequency Spectrum
(unfolded)

a common way to summarize allele frequencies

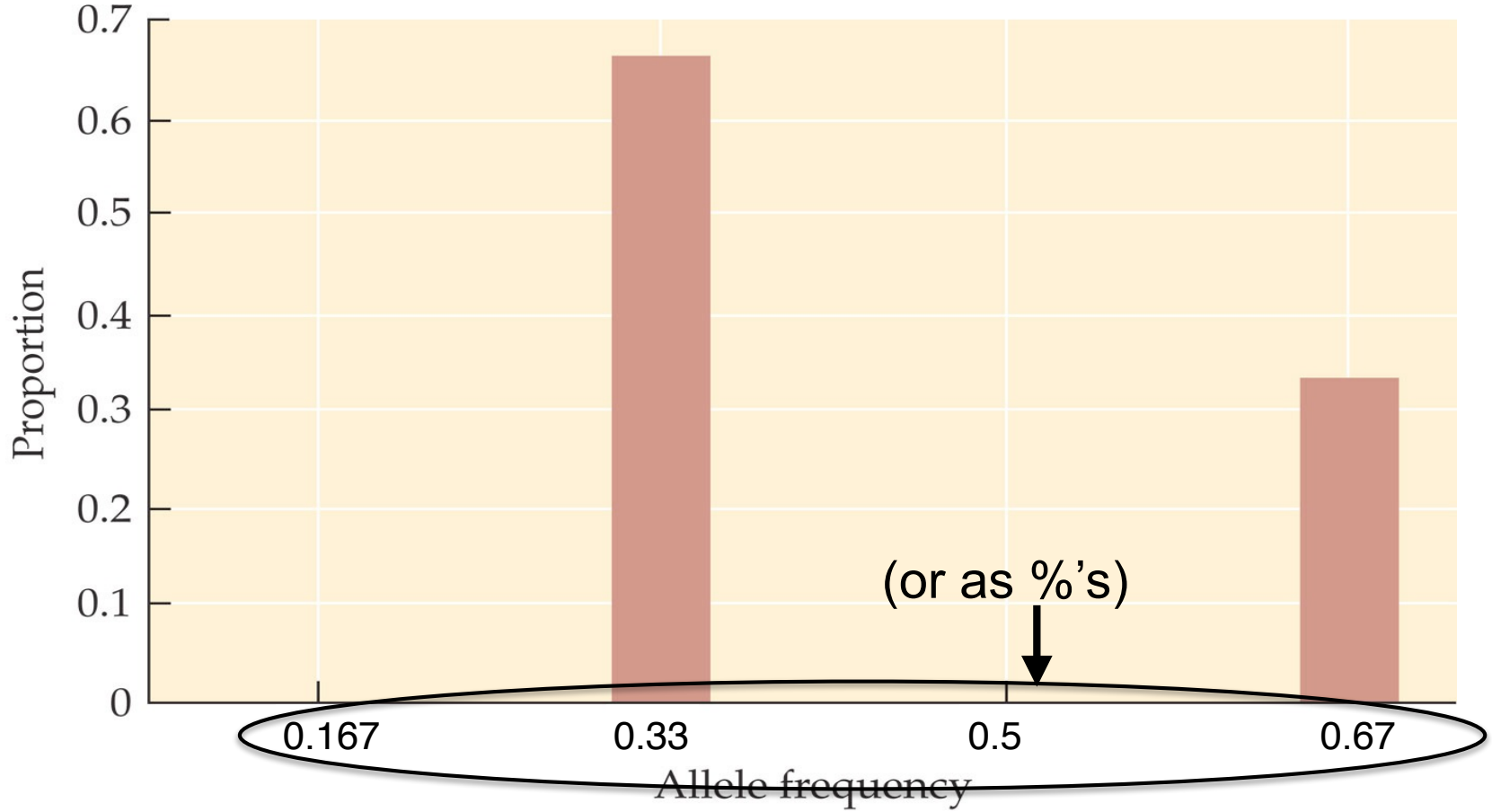


INTRODUCTION TO POPULATION GENETICS, Figure 3.8

the polarity (direction of mutations) matters (unfolded)

Site frequency Spectrum
(unfolded)

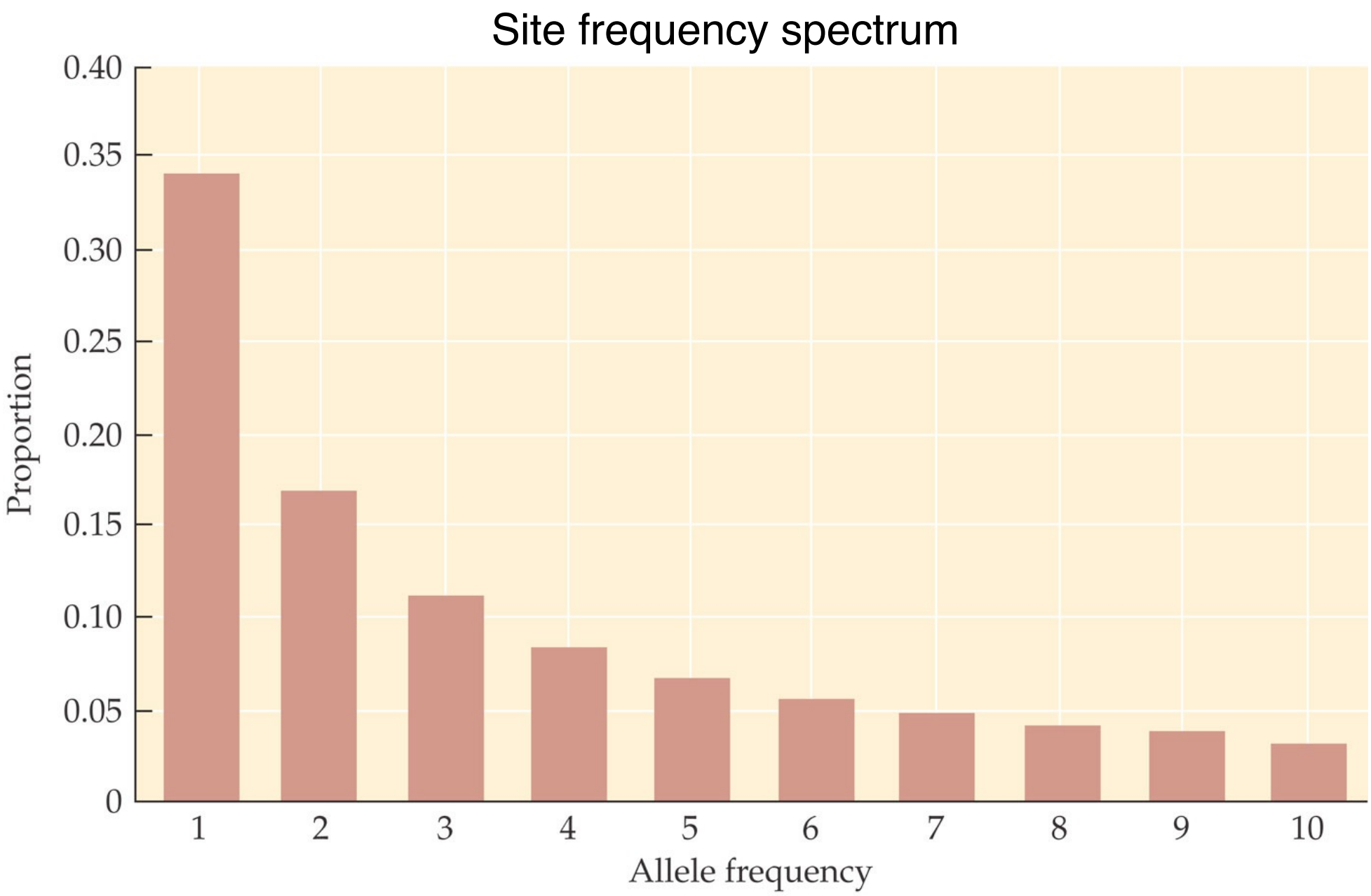
a common way to summarize allele frequencies



INTRODUCTION TO POPULATION GENETICS, Figure 3.8

the polarity (direction of mutations) matters (unfolded)

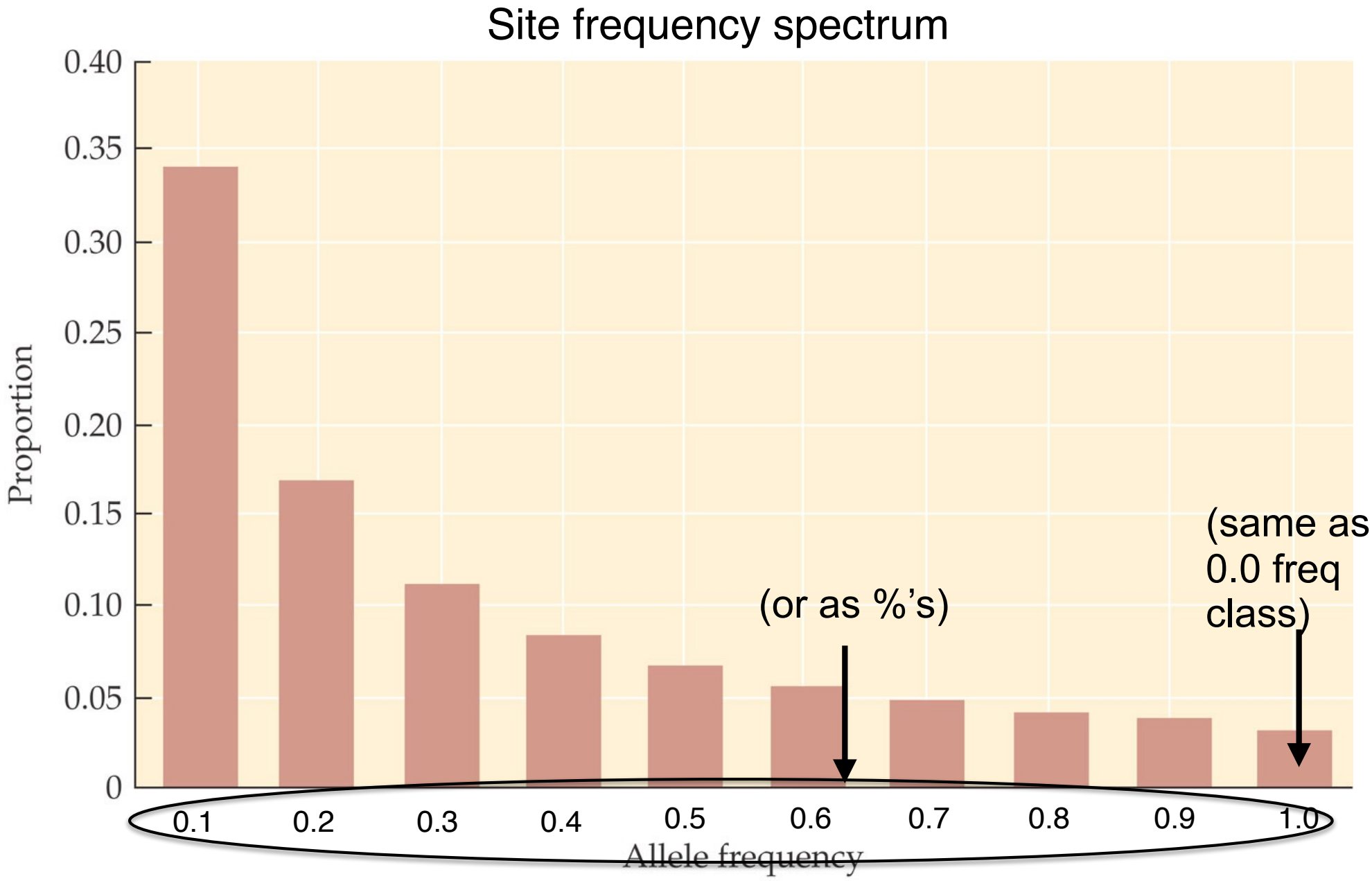
Figure 3.9 The expected site frequency spectrum (SFS) for a sample of n = 10 haploid individuals under the standard neutral coalescence model with infinite sites mutation



INTRODUCTION TO POPULATION GENETICS, Figure 3.9

typical pattern for expanding population

Figure 3.9 The expected site frequency spectrum (SFS) for a sample of n = 10 haploid individuals under the standard neutral coalescence model with infinite sites mutation



INTRODUCTION TO POPULATION GENETICS, Figure 3.9

typical pattern for expanding population

Some SLiM-based simulation exercises

https://github.com/petrelharp/popbio/blob/master/notebooks/slim_intro.ipynb

SLiM Life Cycle(s)

The sequence of events within one generation in WF models.

1. Execution of `early()` events
2. Generation of offspring:
   - 2.1. Choose source subpop
   - 2.2. Choose parent 1
   - 2.3. Choose parent 2 (`mateChoice()` callbacks)
   - 2.4. Generate the offspring (`recombination()` callbacks)
   - 2.5. Suppress/modify child (`modifyChild()` callbacks)
3. Removal of fixed mutations
4. Offspring become parents
5. Execution of `late()` events
6. Fitness value recalculation using `fitness()` callbacks
7. Generation count increment

The sequence of events within one generation in nonWF models.

1. Generation of offspring:
   - 1.1. Call `reproduction()` callbacks for individuals
   - 1.2. The callback(s) make calls requesting offspring
   - 1.3. Generate the offspring (`recombination()` callbacks)
   - 1.4. Suppress/modify child (`modifyChild()` callbacks)
2. Execution of `early()` events
3. Fitness value recalculation using `fitness()` callbacks
4. Viability/survival selection
5. Removal of fixed mutations
6. Execution of `late()` events
7. Generation count increment, individual age increments

A nonWF model

```
basic_nonWF = """
initialize()
{
    // since the model will be non-Wright-Fisher, we need to say so
    initializeSLiMModelType("nonWF");
    // genome is the same as above
    initializeMutationRate(1e-7);
    initializeMutationType("m1", 0.5, "f", 0.0);
    initializeGenomicElementType("g1", m1, 1.0);
    initializeGenomicElement(g1, 0, 99999);
    initializeRecombinationRate(1e-8);
}

reproduction() {
    subpop.addCrossed(individual, subpop.sampleIndividuals(1));
}

1 early() {
    sim.addSubpop("p1", 10);
}

1: {
    catn(sim.generation + " : population size : " + p1.individualCount);
}

10 {
    sim.simulationFinished();
}
"""
out, logfile = slim_script(basic_nonWF, "basic_nonWF", quiet=False)
```

SLiM's Fitness

We need population regulation!

The easiest way is by the `fitnessScaling` attribute of individuals.

In nonWF SLiM models, the "fitness" of an individual is the probability of survival until the next time step.

This is computed by multiplying the fitness calculated from the genome (since all our mutations are neutral, this is `1.0` for everyone) by the `fitnessScaling` attributes of the individual and her subpopulation.

Concretely, we'll fix a "carrying capacity" K, and say that the probability of survival, per individual, when there are N individuals, is K/N.

What is the stable population size of this model? Predict before you run it.

Solution: if we begin with N individuals, then first the population doubles to 2N (everyone reproduces, since by default they are hermaphrodite), then some die, leaving a proportion K/(2N) of the 2N individuals, so we are left with K individuals, on average.

```
basic_nonWF = """
initialize()
{
    // since the model will be non-Wright-Fisher, we need to say so
    initializeSLiMModelType("nonWF");
    // genome is the same as above
    initializeMutationRate(1e-7);
    initializeMutationType("m1", 0.5, "f", 0.0);
    initializeGenomicElementType("g1", m1, 1.0);
    initializeGenomicElement(g1, 0, 99999);
    initializeRecombinationRate(1e-8);

    // this will be the carrying capacity
    defineConstant("K", 1000);
}

reproduction() {
    subpop.addCrossed(individual, subpop.sampleIndividuals(1));
}

1 early() {
    sim.addSubpop("p1", 1000);
}

early() {
    p1.fitnessScaling = K / p1.individualCount;
}

1: early() {
    cat(sim.generation + " : early population size:" + p1.individualCount);
}

1: late() {
    catn( " : late population size:" + p1.individualCount);
}

100 {
    sim.simulationFinished();
}
"""
out, logfile = slim_script(basic_nonWF, "basic_nonWF")
```

Simulation assignment part 1

Option 1 - SLIMgui

Read sections 1-4 of the SLiM Manual Simulate a panmictic population and output the Site frequency spectrum (/ Simulation/Graph mutation frequency Spectrum)) from a population with 2 different combinations of Ne (population size) and mutation rate

Option 2 - momi2 using msprime

do the above with momi2 using msprime backend. momi2 installation instructions can be found at https://radcamp.github.io/IBS2019/07_momi2_API.html

Option 3 - use jerome's msprime tutorial to simulate the SFS with 2 different combinations of Ne (population size) and mutation rate.

https://github.com/DRL/SMBE-SGE-2019

hit the binder button at the bottom to start virtual notebook sessions. Check out session 1 - "2.Introduction_to_msprime.ipynb"

goto to section 2.4 mutation

Instead of the site frequency spectrum, output the gene tree graphs. Hint: add ", Ne = #" (# is some number of your choice)

Simulation assignment part 2 (from last week):

a. Add a beneficial mutation type to SLiM recipe 4 “Neutral simulations in a panmictic population”, sampling 10 genomes from the population at the end
b. How do the two mutation types behave as a function of population size or mutation rate?

Next Week

Present simulation assignment (randomly pick 2 people …. rest turn in code)

Lecture Topic: Coalescent (Chapter 6)