Notes for MCTP Week 2, 2014

Lecture 1: Biological background

Evolutionary biology and population genetics are highly interdisciplinary areas of research, with many contributions being made from mathematics, statistics, and computer science, as well as biology, genetics, and anthropology. Evolutionary biology and genetics both rely heavily on mathematical and probabilistic models. As a result, this workshop will have three main threads: biological background and modeling, probability, and computing.

Models provide idealized assumptions that can be formulated mathematically and used to make predictions about patterns in data. This idea is common in physics. For example, you might assume that there is no wind resistance to simplify calculations of trajectories, or assume that the Earth is perfectly spherical in calculating its gravitational pull on the moon. Similar assumptions are made in biological models. For example, one might assume that a population is infinitely large in order to predict changes in gene frequencies over several generations. Mathematicians and statisticians make contributions to evolutionary biology by helping to formulate such models and to analyze properties of these models and determine how data can be used to fit such models. Computer scientists develop algorithms for handling biological data efficiently. The following questions are typical and each has a different disciplinary emphasis:

1. Can alternative models lead to the same predictions? (mathematics)
2. If a model depends on certain parameters, such as time or population size, what happens when these parameters approach limiting values such as 0 or infinity? (mathematics)
3. Can data be used to estimate reasonable values of the model? (statistics)
4. If so, what functions of the data make good estimates? (statistics)
5. What is an efficient way to summarize data in order to make accurate predictions? (computer science)
6. How can you compute the smallest number of events (e.g., mutations) needed to explain the variation in a data set? (computer science)

Although I have indicated these questions as being typical of different disciplines, people working in different areas often collaborate on the same project, or an individual from one background might work on a question typical of another discipline. For example, many biologists, mathematicians, and statisticians have written computer programs and developed novel algortihms for analyzing data. In many cases, reserachers are more interested in answering questions that are scientifically interesting and are not necessarily concerned with which discipline their research falls in. For this reason, you also see mathematicians, statisticians, and computer scientists publishing in biology journals, statisticians and biologists publishing in mathematical biology journals, and so on.

Phylogenetics and Population Genetics

The subject of this workshop, the coalescent model, is also at the boundary of two areas within biology: phylogenetics and population genetics. For much of their history, these fields have developed fairly independently and have tended to ask different questions:

PHYLOGENETICS — understanding relationships (ancestry) **between** species.

Typical questions in phylogenetics include:

- Are humans more closley related to chimpanzees or gorillas?

- Given DNA sequence data from several species, what evolutionary tree best explains the data?
- How long ago was the most recent common ancestor for a set of species?

POPULATION GENETICS — understanding genetic variation **within** species or populations.

Some typical questions of population genetics include:

- How genetically variable is a population?
- What is the probability that a new mutation will spread throughout a population?
- How quickly do mutations arise?
- How long ago was the most recent common ancestor for a set of genes sampled from a population?

Phylogenetics and population genetics have traditionally been studied separately. In this course we will bridge the gap between the two areas and discuss how phenomena at the population level (population genetics) can have an impact when studying multiple populations or species (phylogenetics).

We begin with classical population genetics.

Mendelian inheritance

Often in a population, there is more than one version of a gene. Different versions of genes are called **alleles**. For example, the blood type you have (A, B, or O) depends on which alleles you have for that gene. You have two copies of each gene for **autosomal** cells (meaning cells other than egg or sperm). For the ABO system, the rules are:

allele	allele	blood type (phenotype)
A	A	A
A	O	A
A	B	AB
B	B	B
B	O	B
O	O	O

The **phenotype** is the observed blood type, while the combination of alleles, such as AB, BB, or OO, is the **genotype**. Often in genetics, the phenotype is observed, while the genotype cannot be observed directly. In some cases, the genotype can be inferred. For example, if a person has type O blood, then they must have the OO genotype. If a person has type A blood, then they have either the AA or AO phenotype. Note that we do not consider the order of the alleles (AO versus OA), so we usually write the alleles in alphabetical order.

If both copies of the gene have the same allele (e.g. AA or OO), then the genotype is said to be **homozygous**. An allele is **dominant** if the same phenotype is observed for the allele in homozygous and heterozygous genotypes. For example, A is dominant to O because either genotype AO or AA leads to same phenotype. The O allele is **recessive**, meaning that the genotype must be OO in order for the blood type to be type O. If there are two distinct alleles at the same locus (e.g. AO or AB), then the genotype is heterozygous. Because the AB genotype leads to a different phenotype from either the AA or BB genotypes, we say that A and B are **co-dominant**.

Exercise. Is B dominant, co-dominant, or recessive with respect to O?

Gametes, which are egg or sperm cells, have one copy of each gene. In a **zygote** (an egg fertilized by a sperm), there is one copy of each gene gene contributed from each parent. This is Mendel's First Law:

The Law of Segregation: each gamete contains one copy of each gene, inherited from either the mother or the father.

Mendel's Second Law states that alleles of different genes assort independently when gametes are formed:

The Law of Assortment: alleles sort independently in gametes, that is, for genes at two loci, the allele in the first locus (maternally or paternally inherited) is independent of the allele at the second locus (maternally or paternally inherited).

We furthermore assume that if an individual is heterozygous at a locus, then there is 1/2 probability for each of the alleles to be present in a gamete.

According to the second law, if in an autosomal cell with two copies of each gene, the first locus has two alleles, say A_1 and A_2, and the second locus has two alleles, say B_1 and B_2, then in the gametes formed by this individual, the probability that a gamete has B_1 given that the gamete has A_1 is equal to the probability that the gamete has allele B_1. In symbols, $P(B_1 \mid A_1) = P(B_1)$.

The Law of Assortment applies if genes are on different chromosomes or are sufficiently separated on the same chromosome. For loci that are sufficiently close on the same chromosome, alleles are often inherited

together and do not assort independently. The degree to which assortment lacks independence can be used as a measure of how close two loci are. This idea has been used to try to find genes (loci) associated with diseases (more later).

The genetic code and mutations

Point mutations

What makes two alleles different from each other? It is not always clear when two copies of a gene should be counted as different alleles or not. Genes are coded for on part of a chromosome by nucleotide sequences, abbreviated as A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). The A and G nucleotides are called **purines** while C and T are called **pyrimadines**.

Chromosomes have a linear arrangement of millions of nucleotides. Genes occur on chromosomes and are typically a few hundred to several thousand nucleotides long. Different alleles might have slightly different sequences of nucleotides. When genes are "expressed" in the process of transcribing and translating them into proteins, the slightly different sequences of nucleotides might or might not result in detectable differences in the resulting protein produced.

One reason for this is the redundancy in the genetic code. Blocks of three nucleotides get translated into **amino acids**. There are 20 amino acids, but there are $4^3 = 64$ possible sets of three nucleotides. Consequently, sometimes two or more triples of nucleotides code for the same amino acid. An example, is that `CCA`, CCC, `CCG`, and `CCT` all code for proline. Generally, the third nucleotide has the most redundancy. A mutation that doesn't change the amino acid, and therefore the protein that the gene codes for, might not result in detectably different phenotypes.

An example is the Alcohol dehydrogenase (*ADH*) locus in fruit flies. This locus has 768 bases of coding sequence. Below is given part of the sequence, from position 571 to 609, sampled from one fruit fly.

```
Fly 1:  gtg.cac.aag.ttc.aac.tcc.tgg.ttg.gat.gtt.gag.ccc.cag
Fly 1:  Val.His.Lys.Phe.Asn.Ser.Trp.Leu.Asp.Val.Glu.Pro.Gln
```

The same part of the sequences, but taken from different individual fruit flies of the same species is shown below:

```
Fly 2:  gtg.cac.aag.ttc.aac.tcc.tgg.ttg.gat.gtt.gag.cct.cag
Fly 2:  Val.His.Lys.Phe.Asn.Ser.Trp.Leu.Asp.Val.Glu.Pro.Gln

Fly 3:  gtg.cac.acg.ttc.aac.tcc.tgg.ttg.gat.gtt.gag.cct.cag
Fly 3:  Val.His.Thr.Phe.Asn.Ser.Trp.Leu.Asp.Val.Glu.Pro.Gln
```

In this example, the three sequences are different at the nucleotide level, with the second sequence having one nucleotide different from the first sequence at position 606 (second to last amino acid). The two sequences are identical at the amino acid level. Whether these are considered distinct alleles might depend on the study one is performing, depending on whether genetic information is being tracked at the nucleotide level or protein level, for example. The third sequence has the same substitution at position 606 as Fly 2, but additionally has a mutation at position 578 (from `aag` to `acg`) which results in an amino acid change from histidine to threonine.

Exercise. How many possible sequences are there with 768 nucleotides? How many possible amino acid sequences are there with $768/3 = 256$ amino acids? What is the ratio of the number of possible nucleotide sequences to amino acid sequences when there are 768 nucleotides? (Hint. To compute the ratio, use logarithms first to manipulate the ratio so that your calculator (or R) can handle the large exponents.)

We can define a **point mutation** as a change in the DNA at a single nucleotide whether or not it results in an amino acid change. Point substitutions that do not result in an amino acid change are called **synomymous** substitutions. If a point substitution does result in an amino acid change, it is called **nonsynonymous**. Mutations between nucleotides of the same type (purine to purine or pyrimadine to pyrimadine) are called **transitions**, while mutations between nucleotides of different types are called **transversions**. This is indicated below:

mutation	type
A $\leftrightarrow$ G	transition
C $\leftrightarrow$ T	transition
A $\leftrightarrow$ C	transversion
A $\leftrightarrow$ T	transversion
G $\leftrightarrow$ C	transversion
G $\leftrightarrow$ T	transversion

Although there are more ways for transversions to occur than transitions, transitions occur more frequently than transversions. Thus, not all point mutations are equally likely. Models for mutation rates for different types of point mutations play an important role in molecular evolution and phylogenetics. In population genetics, a single parameter μ giving the average probability of a point mutation is often used.

Most of the genome consists of DNA that does not code for protein at all and either may serve a regulatory function, influencing gene expression (e.g., when a cell uses the gene to create proteins), or does not have a known function. In addition to noncoding DNA in between genes on a chromosome, there are chunks of DNA within genes, called **introns** that get spliced out of the gene before it is translated into protein. The portions of a gene that are eventually translated into protein are called **exons**.

Other types of mutations

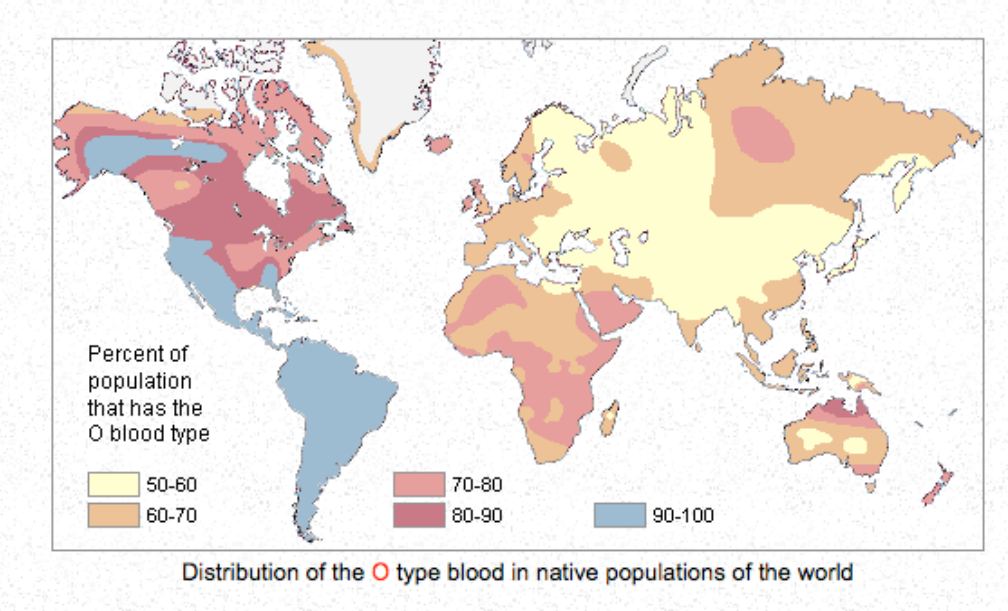
In this course, we will mostly look at point substitutions because they are the easiest to model and are useful for making inferences about ancestry, population sizes, and so forth. However, there are many other types of mutations:

- indels (insertion or deletion)
  - — insertions, e.g. `CCTGTGGCC` $\rightarrow$ `CCTGTGAAAGCC`
  - — deletions, e.g., `CCTGTGAAAGCC` $\rightarrow$ `CCTGTGGCC`
  - — frameshift, occurs when an indel is not a multiple of length 3, resulting in the reading frame for the amino acid sequence not lining up after the indel
- duplications, creating an extra copy of a segment of the gene genome, such as an extra copy of a gene
- translocations, a gene or segment of chromosome to a new location
- inversions, a segment of chromosome gets copied backwards

Polymorphism

Polymorphism means any variable trait, and includes multiple alleles or even single bases that are variable in a population. Single nuceotide polymorphisms, **SNPs**, refer to places in the genome where a single nucleotide has multiple alleles (usually 2). For individuals in a population, there are typically at most two variants at a single nucleotide — this occurs at roughly 1 in 1000 positions in the genome. Since the human genome has over 3 billion base pairs, this means that there are several million SNPs. If there is no variation at a site, then the site is **monomorphic**.

For SNPs with two variants, the more common allele (the one possessed by more than 50% of individuals) is called the **major allele**, while the other is called the **minor allele**. Because there are often just two alleles for a SNP, they are often coded as 0 or 1 rather than using the nucleotide letter.



Distribution of the O type blood in native populations of the world

Patterns of Polymorphism

Population genetics is largely concerned with patterns and dynamics of polymorphism. For example, describing patterns of polymorphism throughout the world for humans (which populations have the most polyphormism), and understanding how and why polymorphism might be maintained in a population (e.g., heterozygote advantage, mutation rates, etc.).

An example is the type O allele. Although in most human populations, this allele occurs more than 50% of the time at this locus, the frequency varies depending on geography.

By tracking allele frequencies, population geneticists try to

- understand patterns of migration and history for populations
- determine whether an allele is under pressure from natural selection
- determine whether populations are at equilibrium or are evolving

We next turn to testing whether the alleles at a locus are under a type of equilibrium, meaning that the allele frequencies are stable over generations.

Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium is a mathematical model that relates genotype frequencies at a locus to allele frequencies and shows that under certain assumptions, equilibrium is reached in one generation. The assumptions used are:

Table 1: Mating outcomes for Hardy-Weinberg Equilibrium

Mating type	offspring	frequency of mating type
$A_1A_1 \times A_1A_1$	A_1A_1	u^2
$A_1A_1 \times A_1A_2$	$\frac{1}{2}A_1A_1 + \frac{1}{2}A_1A_2$	$2uv$
$A_1A_1 \times A_2A_2$	A_1A_2	$2uw$
$A_1A_2 \times A_1A_2$	$\frac{1}{4}A_1A_1 + \frac{1}{2}A_1A_2 + \frac{1}{4}A_2A_2$	v^2
$A_1A_2 \times A_2A_2$	$\frac{1}{2}A_1A_2 + \frac{1}{2}A_2A_2$	$2vw$
$A_2A_2 \times A_2A_2$	A_2A_2	w^2

1. infinite population size
2. discrete generations
3. random mating (for the locus under consideration)
4. no selection (at the locus)
5. no migration
6. no mutation (at the locus)
7. equal genotype frequencies for female and male individuals.

We start assuming that there are two alleles, A_1 and A_2, although the model can be generalized to any number of alleles. As with the ABO system, the order doesn't matter, so A_1A_2 is the same genotype as A_2A_1. When two indivudals mate, they can either both be homozygous, both be heterozygous, or one can be homozygous while the other is heterozygous. We don't distinguish the sex of the parent, that is we consider the **mating type** to be the combination of two genotypes when mating, and do not distinguish which sex has which genotype. In addition, we assume Mendel's First Law, that is each parent contributes a gamete with one allele, chosen at random if the parent is a heterozygote. Letting u, v, and w be the proportion of the population with genotypes A_1A_1, A_1A_2, and A_2A_2, the possible mating types, offspring, and probability of the mating type are:

Given these frequencies of mating types, we can determine the frequencies of the genotypes in the offspring generation. For example, the A_1A_1 offspring can occur when the parental mating type is $A_1A_1 \times A_1A_1$, $A_1A_1 \times A_1A_2$, or $A_1A_2 \times A_1A_2$. Therefore the frequency of A_1A_1, A_1A_2, and A_2A_2 in the offspring is:

$$
\begin{aligned}
P(A_1A_1) &= u^2 + uv + \frac{1}{4}v^2 = \left(u + \frac{1}{2}v\right)^2 \\
P(A_1A_2) &= uv + 2uw + \frac{1}{2}v^2 + vw = 2\left(u + \frac{1}{2}v\right)\left(\frac{1}{2}v + w\right) \\
P(A_2A_2) &= \frac{1}{4}v^2 + vw + w^2 = \left(\frac{1}{2}v + w\right)^2
\end{aligned}
$$

We let $p_1 = u + \frac{1}{2}v$ and $p_2 = \frac{1}{2}v + w$. Then we have the following distribution of genotypes in the offspring generation:

$$
\begin{aligned}
P(A_1A_1) &= p_1^2 \\
P(A_1A_2) &= 2p_1p_2 \\
P(A_2A_2) &= p_2^2
\end{aligned}
$$

After a second round of mating, the frequencies of the genotypes in the generation of grandchildren are

$$\begin{aligned}
P(A_1A_1) &= P(A_1A_1 \times A_1A_1) + \frac{1}{2}P(A_1A_1 \times A_1A_2) + \frac{1}{4}P(A_1A_2 \times A_1A_2) \\
&= p_1^2 \times p_1^2 + \frac{1}{2}2(p_1^2 \times 2p_1p_2) + \frac{1}{4}(2p_1p_2 \times 2p_1p_2) \\
&= p_1^4 + 2p_1^3p_2 + p_1^2p_2^2 \\
&= p_1^2(p_1^2 + 2p_1p_2 + p_2^2) \\
&= p_1^2(p_1 + p_2)^2 \\
&= p_1^2 \\
P(A_2A_2) &= P(A_2A_2 \times A_2A_2) + \frac{1}{2}P(A_1A_2 \times A_2A_2) + \frac{1}{4}P(A_1A_2 \times A_1A_2) \\
&= p_2^2 \times p_2^2 + \frac{1}{2}2(p_2^2 \times 2p_1p_2) + \frac{1}{4}(2p_1p_2 \times 2p_1p_2) \\
&= p_2^4 + 2p_1p_2^3 + p_1^2p_2^2 \\
&= p_2^2(p_2^2 + 2p_1p_2 + p_1)^2 \\
&= p_2^2(p_1 + p_2)^2 \\
&= p_2^2 \\
P(A_1A_2) &= 1 - P(A_1A_1) - P(A_2A_2) = 2p_1p_2.
\end{aligned}$$

Thus, given an arbitrary distribution of parental genotypes, u, v, and w, the children and grandchildren (and subsequent generations), have the same distribution of genotypes. So equilibrium is obtained in a single generation.

Exercise. How many mating types are there if there are three alleles, A_1, A_2, and A_3 at a locus?

Exercise. A simpler way of thinking about this is that if alleles A_1 and A_2 have frequencies p and $q = 1 - p$, and the alleles carried by an individual are independent, then the genotype frequencies can be obtained by expanding $(p + q)^2$. Using this simpler approach, suppose a locus has three alleles, A_1, A_2, and A_3, with genotype frequencies p, q, and r, respectively. What are the Hardy-Weinberg Equilibrium genotype frequencies? How would you generalize this when there are n alleles at a locus?

Testing for Hardy-Weinberg Equilibrium

The assumptions for HWE might seem a bit stringent, however, the model is useful because it can serve as a null hypothesis from which to make predictions about genotype frequencies from allele frequencies. Deviations from HWE give evidence that one of the assumptions is false, such as random mating or absence of selection, although it might be difficult to determine which assumptions are false.

There are many possible tests for HWE, including χ^2, likelihood ratio, and exact multinomial tests. The χ^2 is easiest to use and is a large sample test. It might be less powerful than other methods when sample sizes are small and there is a low expected count for one of the genotypes.

Example To illustrate the use of the χ^2 test for HWE, consider the two-allele MN blood group. Here there are two alleles, M and N, which are codominant. In a sample of 747 individuals from Iceland, numbers of each genotype were tabulated (see Table 2).

The allele frequency for M is found by using $p_{\mathrm{M}} = P(\mathrm{MM}) + \frac{1}{2}P(\mathrm{MN})$, where $P(\mathrm{MM})$ are proportions of the MM and MN genotypes found in the sample. Thus the relative frequency of the M allele in the sample is (233/747)+(1/2)(385/747). The frequency of N is found similarly or by subtracting the frequency of M from 1.

The expected counts are taken by multiplying the expected genotype frequency by the sample size. So $05696^2 \times 747 = 242.36$, and so forth. The χ^2 value is found using the usual formula:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Although there are three terms in the sum, and three cells in the table, there is 1 degree of freedom for the χ^2 test. We use the number of classes minus the number of independently estimated allele frequencies minus 1: 3 - 1 - 1 = 1. For one degree of freedom, the $\alpha = 0.05$-level test for a χ^2 is 3.84. We see that the data for the Iceland sample is well within Hardy-Weinberg expectations.

To quantify the deviation from Hardy-Weinberg expectation, we can define a parameter $D_A = P(A_1A_1) - [P(A_1)]^2$ and test the null hypothesis that $D_A = 0$.

Exercise. The alkaline phophatase locus has three three alleles: S, F, and I. From a sample of 332 English people, the following genotype frequencies were observed.

Genotype	Number
SS	141
SF	111
FF	28
SI	32
FI	15
II	5
Total	332

Find the expected number of each genotype under HWE and conduct a χ^2 goodness-of-fit test of HWE. What degrees of freedom should you use?

Exercise. Suppose two populations both in Hardy-Weinberg Equilibrium for a pair of alleles A_1 and A_2 at a single locus. However, there are different allele frequencies in the two populations. In population 1, the allele frequency for A_1 is p, and the allele frequency for A_2 is $q = 1 - p$. For population 2, the allele frequency for A_1 is P, and the allele frequency for A_2 is $Q = 1 - P$, where $p \neq P$. Suppose the populations are mixed in a new population, such that the probabilities that a randomly chosen individual in the new population comes from population 1 and 2 is m and $1 - m$, respectively.

Show that the mixed population has a deficiency of heterozygotes but achieves HWE in one generation with allele proportions $P(A_1) = mp + (1 - m)P$ and $P(A_2) = mq + (1 - m)Q$.

Allele frequencies when there are different frequencies for females and males: X chromosome

Assume that there are the same number of females and males in a population. This means that there are twice as many X chromosomes in females as there are in males. Let p_1 be the frequency of allele A_1 in the population, and let p_f be the probability that a randomly selected chromosome from the population

Table 2: Example for MN blood group

Genotypes				allele frequencies		
	MM	MN	NN	Total	M	N
Observed	233	385	129	747	0.5696	0.4304
Expected	242.36	366.26	138.38	747		
$\chi^2_1 = 1.96$						

of females has an A_1 allele. Let p_m be the probability that a randomly selected chromosome from the population of males has an A_1 allele. The probability that a randomly selected chromosome has an A_1 allele is:

$$\begin{aligned} P(A_1) &= P(A_1|\text{ female })P(\text{ female }) + P(A_1|\text{ male })P(\text{ male }) \\ &= \frac{2}{3}p_f + \frac{1}{3}p_m \end{aligned}$$

What happens if the initial frequencies of A_1 in females and males are p_f^0 and p_m^0 in generation 0, respectively?

In generation 1, each male gets one copy of either A_1 or not A_1 from its mother, so the probability that a male in generation 1 is carrying A_1 is p_f^0. More generally, in generation i, the probability that a male is carrying an A_1 is p_f^{i-1}. For females, a randomly chosen allele in generation i came from her mother in generation $i-1$ with probability 1/2 or her father in generation $i-1$ with probability 1/2. Therefore the probability that an allele chosen at random from a female in generation i is A_1 is $(p_f^{i-1} + p_m^{i-1})/2$. This gives us a recurrence:

$$\begin{aligned} p_f^i &= (p_f^{i-1} + p_m^{i-1})/2 \\ p_m^i &= p_f^{i-1} \\ p_f^i - p_m^i &= (1/2)(p_f^{i-1} - p_m^{i-1}) \end{aligned}$$

The last recurrence in particular means that $p_f^n - p_m^n = (1/2)^n(p_f^0 - p_m^0) \to 0$. In other words, the allele frequencies in men and women are approaching the same value, which is p_1. Example, let $p_f^0 = 0.1$ and let $p_m^0 = 0.5$. Then

i	p_f^i	p_m^i	difference
0	0.1	0.5	-0.4
1	0.3	0.1	0.2
2	0.2	0.3	-0.1
3	0.25	0.2	0.05
4	0.225	0.25	-0.025
5	0.2375	0.225	0.0125

Finite Populations

We saw that under HWE, heterozygosity is stable. That is, the proportion of heterozygotes for allele A_iA_j, $i \neq j$ is expected to remain at $2p_ip_j$. However, HW assumes an infinite population. What happens in a finite population?

Let $\mathcal{G}$ be the probability that two loci are identical by state (IBS) when drawn from the population randomly without replacement (so two copies of the same locus on different chromosomes one individual could be sampled, but not from the same chromosome). After one round of random mating, two randomly sampled loci might be IBS either because the had the same parent allele in the previous generation (with probability $1/(2N)$, or because they were descended from different gene copies that were IBS (with probability $(1 - 1/(2N))\mathcal{G}$. Therefore

$$\mathcal{G}' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}$$

We also have

$$\begin{aligned}1-\mathcal{G}' &= 1-\frac{1}{2N}-\left(1-\frac{1}{2N}\right)\mathcal{G} \\ &= \left(1-\frac{1}{2N}\right)(1-\mathcal{G}) \\ &= \left(1-\frac{1}{2N}\right)\mathcal{H}\end{aligned}$$

where $\mathcal{H}$ is a finite-sample version of heterozygosity.

This leads to an expression for the decay of heterozygosity in finite populations:

$$\mathcal{H} = \mathcal{H}_0\left(1-\frac{1}{2N}\right)^t$$

where t is the number of generations and $\mathcal{H}_0$ is the starting heterozygosity. Thus, heterozygosity decreases ex ponentially over time.

The "half-life" of heterozygosity can be found by solving

$$\begin{aligned}\frac{1}{2}\mathcal{H}_0 &= \mathcal{H}_0\left(1-\frac{1}{2N}\right)^t \\ \Rightarrow \log(1/2) &= t\log\left(1-\frac{1}{2}\right) \\ \Rightarrow t &= -\frac{\log 2}{\log\left(1-\frac{1}{2N}\right)} \\ &\approx 2N\log(2) \approx 1.4N\end{aligned}$$

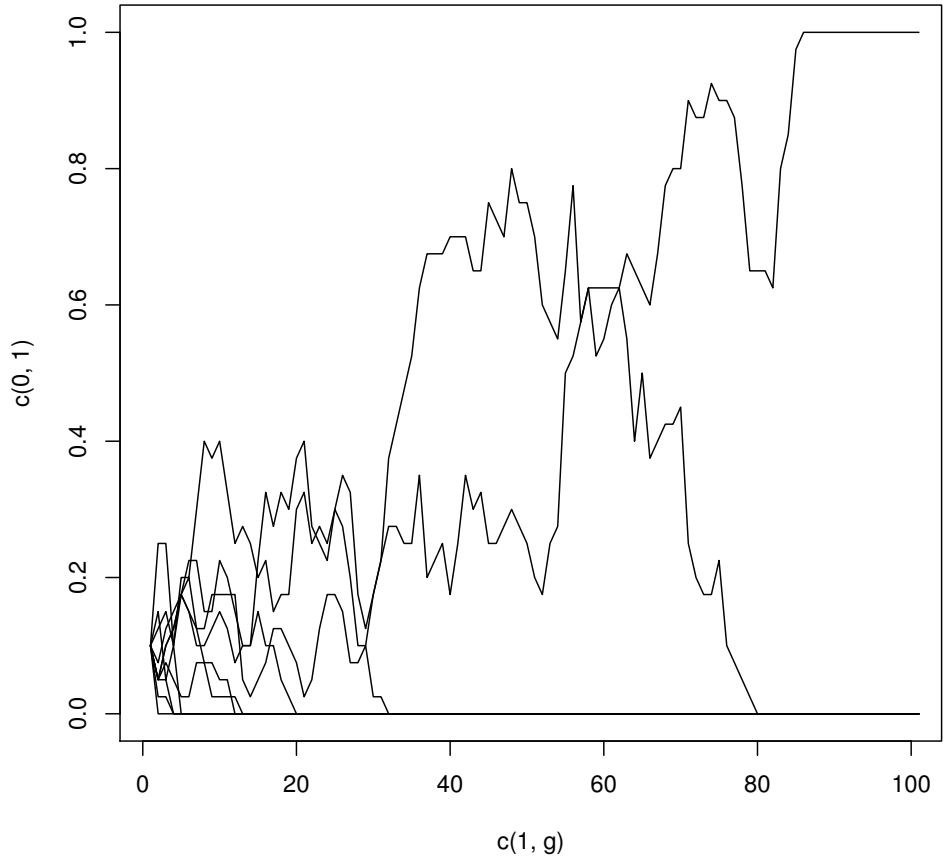
For large populations, it can therefore take a long time for heterozygosity to decrease by 1/2. For example, in a population with 50,000 individuals and a generation time of 20 years, it would take approximately 1.4 million years.

We can model genetic drift in finite populations using a binomial model. We assume that there are $2N$ copies of a locus wit h say, two alleles A_1 and A_2. Because the population is finite rather than infinite, we consider counts for the two alleles rather than proportions. Suppose there are i copies of allele A_1 and $2N-i$ copies of allele A_2. In the model, we assume that each allele in generation $m+1$ is randomly sampled from the previous generation. The number of copies of A_1 in generation $m+1$ is therefore a binomial random variable X with distribution

$$X \sim Bin\left(2N, \frac{i}{2N}\right)$$

We can simulate this process over many generations to observe the long term behaviour in allele frequencies, and the probability that an allele fixates (becomes the only variant in the population).

An example of how to do this is given in the figure.




```
p <- .9  #proportion of A allele
N <- 20  #2N is number of allele copies
g <- 100 #number of generations
I <- 10 #number of iterations

plot(c(1,g),c(0,1),type=''n'')

for(i in 1:I) {

  y1 <- rep(0, round(2*N*p))  # number of copies of allele 0
  y2 <- rep(1, 2*N-round(2*N*p)) # number of copies of allele 1
  x <- c(y1,y2)
  ave <- mean(x)

  for(j in 1:g) {
  x <- sample(x,replace=T)
  ave <- c(ave,mean(x))
  }

  points(ave,type=''l'')
}
```

Wright-Fisher derivation of the coalescent

The coalescent is obtained by considering ancestry in the Wright-Fisher model as the population size N goes to infinity, but we consider the sample size to be small compared to N. Here the sample is a subset of genes (rather than individuals) in the population. The assumptions for the Wright-Fisher are similar to Hardy-Weinberg Equilibrium

1. Generations are discrete
2. Population size is finite and fixed in each generation
3. Ancestry is random, with each gene copy having a random parent in the previous generation
4. There is no mutation
5. There is no selection
6. Genes are independent
7. Generations are independent

In the model, we are not concerned with alleles — different versions of a gene, only with the ancestry of the gene.

Of particular interest is the genealogical tree structure that this process produces when we only pay attention to the history of a sample backward in time. This removes the complexity of keeping track of an entire population and introduces enormous computational advantages over forward-time approaches. Particular questions we are interested in are

1. What is the waiting time until two lineages have a common ancestor?
2. What is the probability that more than two lineages have the same parent in the previous generation (a 3-way merger)?
3. What is the waiting time until the next coalescent event when there are n genes sampled from a population of N genes?
4. What is the waiting time until the Most Recent Common Ancestor of an entire sample of n genes?

The coalescent is only concerned with ancestry, not with mutation. However, once the probability distribution of coalescent trees is understood, we can couple the coalescent with models of mutation and derive other quantities of interest such as the expected number of mutations on a tree, and therefore in a sample of n lineages.

Simulations of mutations are also typically done in two steps: (1) first simulate a tree backwards in time without mutation, (2) then simulate mutations forwards in time at different points on the tree (starting with some ancestral state).

First, we try to understand some of the questions related to waiting times in the Wright-Fisher model with a finite population. The coalescent is derived by taking limits as the population size goes to infinity.

Question 1. Given a sample of two genes from the population, what is the distribution of time until they coalesce?

To answer this, note that the probability that the parent of gene 1 is parent 1 is $1/N$. The probability that the parent of gene 2 is parent 1 is also $1/N$. Since these probabilities are independent, the probability

that genes 1 and 2 both have parent 1 is $1/N^2$. Let E_i be the event that genes 1 and 2 both have parent i. Then $P(E_i) = 1/N^2$. Therefore the probability that genes 1 and 2 have the same parent is

$$P\left(\cup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) = \sum_{i=1}^{n} (1/N^2) = 1/N$$

Another way of looking at this probability is that it does matter what parent gene 1 has. Suppose gene 1 has parent i. The probability that gene 2 also has parent i is $1/N$.

Now, what holds for genes 1 and 2 also holds for any genes i, and j, where $i \neq j$ since all the genes have the same distribution. Thus the probability that any two genes have the same parent in the previous generation is $1/N$.

What is the probability that two genes have different parents in generations $1, 2, 3, \ldots, t-1$, but the same parent in generation t? (We'll let the present generation, from which the genes were sampled, be called generation 0.) The probability that two particularly genes fail to have the same parent in the previous geeneration is $1 - 1/N$. Since this occurs independently in $t-1$ generations, the probability is $(1-1/N)^{t-1}$ that the genes fail to coalesce in the first $t-1$ generations. There is a probability of $1/N$ that they coalesce in the tth generation. Thus the probability of waiting exactly t generations is exactly

$$P(T = t) = (1 - 1/N)^{t-1} t, \quad t = 1, 2, \ldots$$

This distribution is called the Geometric distribution (note the similarity to geometric series). On average, it takes N generations for two lineages to coalesce.

Question 2. The second question asks what is the probability of a 3-way merger in one generation. We might also be interested in the probability that two pairs of lineages coalesce simultaneously. Part of the coalescent derivation is to show that as $N \to \infty$, the most likely events are that either no coalescence occurs, or one pair of lineages coalesce — it is very unlikely for more than two lineages to coalesce in one generation for large N.

To show this, we let $G_{i,j}$ be the probability that i gene copies have exactly j ancestors in the previous generation. No coalescence occurs if $j = i$, and exactly one pair coalesces if $j = i - 1$, so are primarily interested in the probabilities $G_{i,i}$ and $G_{i,i-1}$. For $G_{i,i}$, this is the probability that the i genes in the sample have i distinct ancestors. The probability of this occurring is

$$G_{i,i} = \frac{N \cdot (N-1) \cdot (N-2) \cdots (N-i+1)}{N^i} = 1 \cdot (1 - 1/N) \cdot (1 - 2/N) \cdots (1 - (i-1)/N)$$

The denominator is obtained by each of the i genes having N choices for its parent. The numerator counts the number of ways for the parents to be distinct, with N choices for the first gene, $N-1$ choices for the second gene (since the parent of the first gene is not available), and so on.

For $G_{i,i-1}$, the i lineages in the sample have $i-1$ distinct ancestors. This is like putting i balls into $i-1$ boxes, a situation in combinatorics called "The pigeon-hole principle". In this situation, one of the boxes must have two balls. In the genetics example, one of the ancestors must have two descendants, corresponding to a coalescence. To count the number of ways this could occur, we can choose two of the i lineages to coalesce. There are $\binom{i}{2} = i(i-1)/2$ choices for which two coalesce. There are then $N \cdot (N-1) \cdots (N-(i-1)+1)$ ways to make the $i-1$ ancestors distinct. And again there are N^i equally likely choices for the i lineages. Thus

$$G_{i,i-1} = \frac{\binom{i}{2} N \cdot (N-1) \cdots (N-i+2)}{N^i} = \frac{\binom{i}{2}}{N} \cdot (1 - 1/N) \cdot (1 - 2/N) \cdots (1 - (i-2)/N)$$

Now, what we are interested in $G_{i,j}$ large N.

If we multiply out $G_{i,i}$, we get

$$G_{i,i} = 1 - \frac{\sum_{j=1}^{i-1} j}{N} + O(1/N^2) = 1 - \frac{\binom{i}{2}}{N} + O(1/N^2)$$

Similarly,

$$G_{i,i-1} = \frac{\binom{i}{2}}{N} + O(1/N^2)$$

Note that for large N, $G_{i,i} + G_{i,i-1} \approx 1$, so for large N, we don't have to worry about multiple mergers or simultaneous coalescences.

Therefore the probability that we wait more than t generations for a coalescence is

$$G_{i,i}^t = \left[1 - \frac{\binom{i}{2}}{N} + O(1/N^2)\right]^t \approx \left[1 - \frac{\binom{i}{2}}{N}\right]^t = \left[1 - \frac{\binom{i}{2}}{N}\right]^{N \cdot t/N} = \left(\left[1 - \frac{\binom{i}{2}}{N}\right]^N\right)^{t/N}$$

Now, as $N \to \infty$ and t/N stays moderate, the term on the right approaches $e^{-\binom{i}{2}t/N}$ Similarly, the probability that there is one coalescence within t generations approaches $1 - e^{-\binom{i}{2}t/N}$.

Time here is measured with t in generations, and time scaled by t/N is referred to as coalescent units.

The take home message is that waiting time to the next coalescent event, going backwards in time, is an exponential random variable with rate parameter $\binom{i}{2}$, where i is the number of lineages in the sample. From this idea, many other quantities can be derived. For example, the waiting time until the most recent common ancestor of the sample is the sum of the waiting times when there are i lineages, $i-1$ lineages, ..., 2 lineages. Let T_i be the waiting time when there are i lineages and let $T = \sum_{k=2}^{i} T_i$. Then

$$\begin{aligned}
E[T] &= E\left[\sum_{k=2}^{i} T_k\right] \\
&= \sum_{k=2}^{i} E[T_k] \\
&= \sum_{k=2}^{i} \frac{1}{\binom{k}{2}} \\
&= \sum_{k=2}^{i} \frac{2}{k(k-1)} \\
&= 2\sum_{k=2}^{i} \frac{1}{k-1} - \frac{1}{k} \\
&= 2\left[1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{i-1} - \frac{1}{i}\right] \\
&= 2\left[1 - \frac{1}{i}\right] = \frac{2(i-1)}{i}
\end{aligned}$$

In the limit as $i \to \infty$, $E[T] \to 2$, which might seem remarkable. As the sample size gets larger and larger, the tree doesn't explode in terms of time to the MRCA; instead the expected height of the tree approaches the finite value of 2 coalescent units. The property that the expected time to the MRCA doesn't approaches a finite value (rather than infinity) when the sample size approaches infinity is known as the "coming down from infinity" property, and it is partly a result of the rapid (quadratic) increase in rates with the sample size.

Another quantity of interest is the sum of the lengths of all branches. In each epoch (time between having i and $i-1$ lineages, there are i branches of length T_i. Thus the expected total length of the branches of the

tree is

$$\begin{aligned} E[T_{tot}] &= \sum_{k=2}^{i} E[kT_k] \\ &= \sum_{k=2}^{i} \frac{k}{\binom{k}{2}} \\ &= \sum_{k=2}^{i} \frac{2k}{k(k-1)} \\ &= 2\sum_{k=2}^{i} \frac{1}{k-1} \\ &= 2\sum_{k=1}^{i} \frac{1}{k} \end{aligned}$$

Therefore $E[T_{tot}]$ does approach infinity as the sample size goes to infinity; however it does so slowly with only logarithmic growth, and $E[T_{tot}] \to 2(\log(i)) + \gamma$ for large i, where γ is Euler's constant.

Another type of property we are interested in is the probability that i lineages have exactly j ancestors as a function of continuously varying time t. This generalizes the probability $G_{i,j}$ discussed earlier which gives this probability in the special case that time is 1 generation. Let the time-dependent probability that i lineages have exactly j ancestors as a function of continuously varying time t be denoted $g_{i,j}(t)$. From the exponential distribution, we already know that

$$g_{2,1} = 1 - e^{-t} \quad g_{2,2}(t) = e^{-t}$$

How do we know this? The waiting time to coalescence is exponential with rate $\binom{i}{2}$. When $i = 2$, the waiting time is therefore exponential with rate 1. For an exponential random variable T_2 (recall that T_2 has rate 1, $P(T_2 > t) = e^{-t}$ and $P(T_2 < t) = 1 - e^{-t}$. A waiting time being less than t corresponds to two lineages coalescing into one lineages within time t, so

$$g_{2,1}(t) = P(T_2 < t) = 1 - e^{-t} \quad g_{2,2}(t) = P(T_2 > t) = e^{-t}$$

To derive $g_{3,1}(t)$, we could use the following. The probability that three lineages coalesce into one within time t is $P(T_3 + T_2 < t)$. This probability can be determined in a couple of ways. One way is to use a double integral:

$$
\begin{aligned}
P(T_3 + T_2 < t) &= P(T_3 < t - T_2) \\
&= \int_0^t \int_0^{t-t_2} 3e^{-3t_3} e^{-t_2} \, dt_3 \, dt_2 \\
&= \frac{3}{3} \int_0^t e^{-t_2} (-e^{-3t_3}) \Big|_0^{t-t_2} \, dt_2 \\
&= \int_0^t e^{-t_2} (1 - e^{-3(t-t_2)} \, dt_2 \\
&= \int_0^t e^{-t_2} - e^{-3t+2t_2} \, dt_2 \\
&= \int_0^t e^{-t_2} \, dt_2 - e^{-3t} \int_0^t e^{2t_2} \, dt_2 \\
&= 1 - e^{-t} - e^{-3t} \cdot \frac{1}{2} e^{2t_2} \Big|_0^t \\
&= 1 - e^{-t} - \frac{e^{-3t}(e^{2t} - 1)}{2} \\
&= 1 - \frac{3}{2} e^{-t} + \frac{1}{2} e^{-3t}
\end{aligned}
$$

An easier probability to work out is $g_{3,3}(t) = P(T_3 < t) = e^{-3t}$. Note that for this case, if $T_3 > t$, then there are still three lineages after time t, so we don't need to worry about T_2 at all. Since three lineages coalesce into either 1,2, or 3 lineages, the final case to consider is $g_{3,2}(t)$. This can be obtained using the fact that these three probabilities must sum to 1:

$$g_{3,1}(t) + g_{3,2})(t) + g_{3,3}(t) = 1$$

so that $g_{3,2}$ can be obtained by subtracting the other two probabilities from 1. This yields

$$g_{3,2}(t) = \frac{3}{2} e^{-t} - \frac{3}{2} e^{-3t}$$

It is also possible to derive $g_{3,2}(t)$ directly, using $g_{3,2}(t) = P(T_3 < t < T_3 + T_2)$. Again, this can be solved as a double integral:

$$P(T_3 < t < T_3 + T_2) = \int_0^t \int_{t-t_3}^{\infty} 3e^{-3t_3} e^{-t_2} \, dt_2 \, dt_3$$

What if we wanted to derive $g_{4,1}(t)$? Using the approach here, this would be a triple integral. If we wanted $g_{10,1}(t)$, this would be a 9-dimensional integral. Another approach to work out the distribution of $S_{i,j} = \sum_{k=j}^{i} T_k$ where T_k is exponential with parameter $\binom{k}{2}$. If this is true, then something like

$$g_{9,4}(t) = P(T_2 + T_3 + T_4 + T_5 < t < T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8 + T_9) = P(S_{2,5} < t < S_{2,5} + S_{6,9})$$

could be accomplished using just two independent random variables. The distribution of $S_{j,k}$ has been derived and is known as a hypoexponential random variable, where it is assumed that each of the T_k have different rate parameters (which is true in this problem).

Fortunately, a generally formula for $g_{i,j}(t)$ was derived in 1984 by Tavaré.

The formula assumes that $1 \leq j \leq i$ and $t > 0$.

$$g_{i,j}(t) = \sum_{k=j}^{n} e^{-\binom{j}{2} t/2} \cdot \frac{(2k-1)(-1)^{j-k} j_{(k-1)} n_{[k]}}{j!(k-j)! n_{(k)}}$$

where $n_{(k)} = n(n+1) \cdot (n+k-1)$ and $n_{[k]} = n(n-1) \cdot (n-k+1)$. Some properties of this formula include:

1. For any $i \geq 1$ and any $t > 0$, $\sum_{j=1}^{i} g_{i,j}(t) = 1$
2. For any $i \geq 2$, as $t \to \infty$, $g_{i,1}(t) \to 1$ and $g_{i,j}(t) \to 0$ for $j > 1$
3. For any $i \geq 1$, as $t \to 0$, $g_{i,i}(t) \to 1$ and $g_{i,j}(t) \to 0$ for $j < i$.
4. $\lim_{n\to\infty} g_{i,j}(t) = \sum_{k=j}^{\infty} e^{-\binom{j}{2}t/2} \cdot \frac{(2k-1)(-1)^{j-k} j_{(k-1)}}{j!(k-j)!}$
5. In practice it is difficult to computationally to evaluate $g_{i,j}(t)$ for large i, so $g_{\infty,j}(t)$ is often used. Note that this is the "coming down from infinity" idea again

Gene trees in species trees: Application of $g_{i,j}(t)$ function