G14TBS Part II: Population Genetics

Dr Richard Wilkinson
Room C26, Mathematical Sciences Building
Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

Spring 2015

Preliminaries

These notes form part of the lecture notes for the module G14TBS. This section is on Population Genetics, and contains 14 lectures worth of material. During this section will be doing some problem-based learning, where the emphasis is on you to work with your classmates to generate your own notes. I will be on hand at all times to answer questions and guide the discussions.

Reading List: There are several good introductory books on population genetics, although I found their level of mathematical sophistication either too low or too high for the purposes of this module. The following are the sources I used to put together these lecture notes:

- Gillespie, J. H., *Population genetics: a concise guide*. John Hopkins University Press, Baltimore and London, 2004.
- Hartl, D. L., *A primer of population genetics, 2nd edition*. Sinauer Associates, Inc., Publishers, 1988.
- Ewens, W. J., *Mathematical population genetics: (I) Theoretical Introduction*. Springer, 2000.
- Holsinger, K. E., *Lecture notes in population genetics*. University of Connecticut, 2001-2010. Available online at

  `http://darwin.eeb.uconn.edu/eeb348/lecturenotes/notes.html`.
- Tavaré S., *Ancestral inference in population genetics*. In: Lectures on Probability Theory and Statistics. Ecole d'Etés de Probabilité de Saint-Flour XXXI – 2001. (Ed. Picard J.) Lecture Notes in Mathematics, Springer Verlag, 1837, 1-188, 2004. Available online at

  `http://www.cmb.usc.edu/people/stavare/STpapers-pdf/T04.pdf`

Gillespie, Hartl and Holsinger are written for non-mathematicians and are easiest to follow. Ewens and Tavaré are more technically challenging, but both are written by pioneers in population genetics and are well worth looking through.

Contents

Chapter 1

Everything you wanted to know about genetics, but were afraid to ask

Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

If you know nothing about genetics, read through *DNA from the beginning*, available at http://dnaftb.org/

We will now cover the very basics needed for this course. Genetics essentially separates into two eras

- Classical genetics (Mendelian inheritance)
- Molecular genetics

1.1 Some history

We will describe briefly the development of genetics over the past 150 years, but using modern terminology throughout.

1.1.1 Classical genetics

The study of genetics began with the study of traits and heredity, and it has long been known that traits are passed

from parents to offspring. However, it wasn't until 1865 that Gregor Mendel (an Augustian monk) posited the existence of discrete factors called **genes** inherited from parents. Alternative forms of a trait come from alternative forms of a gene (called **alleles**). By a large number of meticulous experiments with pea plants, Mendel managed to infer the following:

- genes come in pairs
- genes don't blend
- some genes are dominant over others
- each trait is controlled by a pair of genes (he only examined simple traits)

Mendel proposed the following two laws, now known as the Mendelian laws of inheritance:

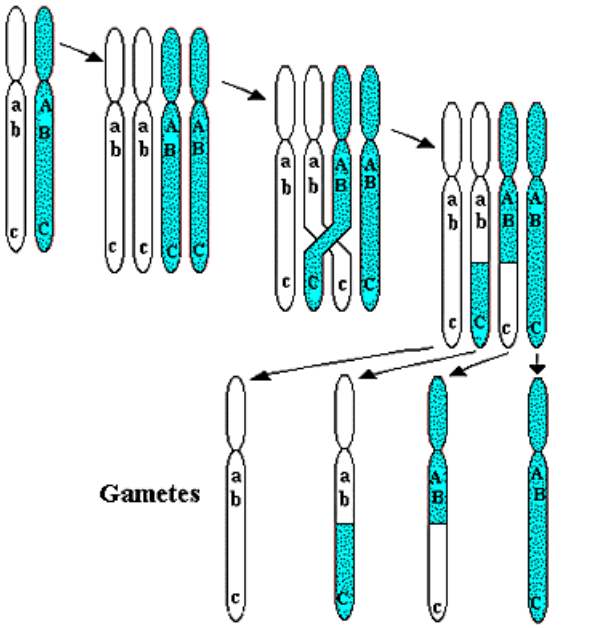
1. **Law of segregation:** when an individual produces gametes (sex cells) the copies of each gene separate so that each gamete receives only one copy (a process known as *meiosis*).
2. **Law of independent assortment:** Alleles of different genes assort independently of one another during gamete formation.

The first law explained how replication works. The second law leads to Mendel's famous 3:1 ratio. Note that the second law is only true for genes that are not linked (see later).

Albinism

Consider a simple trait (determined by a single gene), e.g., albinism. Let A denote the normal gene, and a denote the gene for albinism. We know that A is dominant over a. Everybody carries two copies of this gene. There are three possible **genotypes***: AA, Aa and aa. But there are only two* **phenotypes***: normal or albino.*

We call people with the AA or aa genotypes **homozygous**, *and people with the Aa genotype are called* **heterozygous**



Crossing-over and recombination during meiosis

Figure 1.1: An illustration of meiosis, crossing-over and recombination in the formation of gametes.

(for this gene). Mendelian genetics tells us what happens when different genotypes breed.

For example, if two heterozygous individuals mate, then Mendel's first law says that all three genotypes are possible for their **progeny**. *Mendel also showed that the ratio of normal offspring to non-normal (in this case albino) will be 3:1 (Mendel's experiments were with pea colour, not albinos!). This is because*

$$Aa \times Aa \text{ produces } \begin{cases} AA & \text{w.p. } 1/4 \\ Aa & \text{w.p. } 1/2 \\ aa & \text{w.p. } 1/4 \end{cases}$$

Hence, this breeding pair will produce 3 normal offspring for every albino (on average).

A **Mendelian trait** is one controlled by a single **locus** (gene location on a chromosome) and shows a simple Mendelian inheritance pattern. Other examples include sickle cell anemia and cystic fibrosis.

NB Note the inherently probabilistic nature of Mendelian inheritance.

Quantal nature of genes

Mendel's work was completely ignored until 1900.

At the same time as Mendel was conducting his pea experiments, the structure of the cell was being explored and chromosomes (thin thread-like structures) were discovered. It was suggested that chromosomes carry the information needed for each life form. In the 1870s cell division and chromosome replication (mitosis) was discovered, and it was realised that all life must arise from pre-existing life via replication. It was found that normal body cells (**diploid** cells) have two copies of each chromosome, and that sex cells (gametes) only have one set (haploid cells). Meiosis (formation of gametes) was discovered in the 1890s, and it was shown that meiosis halves the set of chromosomes and randomly assorts **homologous** chromosomes into sex cells (cf. Mendel's second law). See Figure 1.1 for an illustration of meiosis.

It was only once this experimental work was complete, that Mendel's abstract genetic theory was given the physical context it needed.

Genetic Basis of gender

In 1905 the chromosomal basis of gender was discovered. Humans were observed to have 22 homologous pairs of chromosomes, and 1 pair of non-identical chromosomes (in men) (46 chromosomes - 23 pairs in total), and similar phenomena were observed in other species.

The sex chromosome, is XX in women and XY in men, and the X chromosome is much longer than the Y chromosome. Females produce only X gametes, whereas males produce X and Y gametes in equal proportion, explaining the genetic basis of gender.

Morgan ushered in the era of modern genetics when he showed that genes must be physically located on chromosomes. He also showed that some traits tend to occur together, e.g., they are linked. He showed that in *Drosophila melanogaster* (the fruit fly), there were four linked groups of traits, which equals the number of chromosomes pairs in `Drosophila` - this provided further evidence that genes are on chromosomes and began to explain why some traits are linked. They used **linkage** to construct maps of fruit fly chromosomes.

Morgan also observed that linked traits are sometimes separated during meiosis, breaking the laws of Mendelian inheritance. He concluded that how often they separated provided a measure of the relative distance between them on the chromosome. This started the idea that genes are located linearly along a chromosome. Traits determined by genes on the same chromosome tend to be inherited together. However during meiosis, **cross-over** sometimes occurs, and part of each homologous chromosome is passed into the gamete - in this case the linkage is broken. The chance of genes being split in this manner depends on how far apart they are.

Evolution

Darwin published *On the origin of species* in 1859. In it he laid out his theory that evolution occurred by natural selection. In the early 20th century it became clear that mutations in genes are the source of variation and that Mendelian genetics offered a statistical method for analysing inheritance of new mutations. Sadly, these ideas also led to eugenics, which wrongly assumed that complex traits (intelligence, mental illness etc) could be explained by the simple dominant/recessive gene theory of Mendel, leading to the Nazi attempt to purify the German 'race'. We now know that most traits involve many genes, usually in a non-trivial manner.

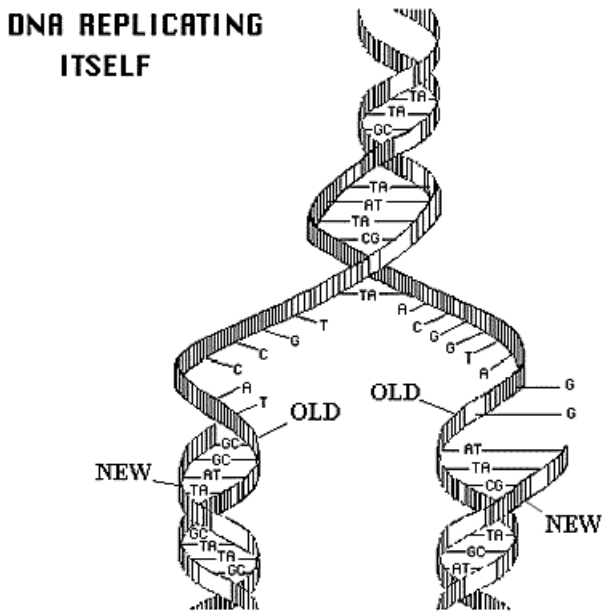


Figure 1.2: A replicating dna molecule.

1.1.2 Molecular Genetics

In the late 19th and early 20th centuries, it was widely believe that proteins were the carriers of genetic information. Proteins are long linear chains of amino acids. There is an amino-acid 'alphabet' of 20 different 'words', and proteins consist of long combinations of different words. We now know that although proteins are the chief actors within a cell, they are not the carriers of genes.

DNA (deoxyribonucleic acid) was discovered as a molecule in 1860, however it wasn't until 1941 that it was discovered that genes are made of DNA. DNA consists of four elements, which are usually labelled A, C, G and T (adenine, cytosine, guanine and thymine). It was initially thought that DNA was just a monotonous sequence of these four elements with no known function.

In 1953 Watson and Crick showed the double helix structure of DNA, which made clear how DNA replicates. In particular, A always pairs with T, and G with C. When the two helical strands separate, each forms a blueprint for a complete dna molecule. See Figure 1.2.

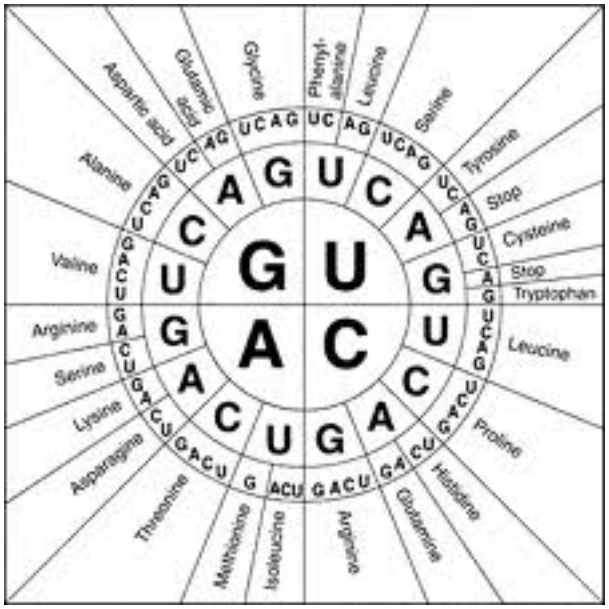


Figure 1.3: Translating codons into amino-acids.

In this course, we will think of dna as a long string of the letters A, C, G and T, which we will call base pairs (bp). Recall that A is always joined to T and C to G and vice versa, so we only need to know the sequence on one strand of the DNA molecule to know the entire sequence. The **'central dogma'** of genetics, is that DNA codes for RNA, which then codes for protein, i.e., genetic information flows from DNA to proteins. The DNA molecule is split into **codons**, which are a sequence of three letters. Each codon codes for a single amino-acid.

There is redundancy in this genetic code, with some codons coding for the same amino-acid, and some being 'start' or 'stop' signals.

Mendel said a gene is a discrete unit of heredity that influences a visible trait. A modern understanding of a gene is that is a discrete sequence of DNA encoding a protein, beginning with a start codon and ending with a stop codon. See Figure 1.3.

1.1.3 Genetic change

Each DNA difference results from a mutation. This can range from

- a single nucleotide change (a SNP - a single nucleotide polymorphism)
- small repeated units in replication
- larger insertions and deletions

and mutations can have external causes, such as radiation, or may be due to errors in replication.

Mutations can

- be beneficial (which might then spread through the population leading to evolution)
- lead to disease
- be neutral.

In humans, the vast majority of mutations are neutral, usually because they occur in non-coding regions. Only 5% of the human genome is thought to code for proteins, the rest being non-coding.

Within humans, genetic variation is usually about 1 in 1000 base-pairs difference.

The human-genome, completed in 2001, successfully sequenced the entire human genome in one individual. Scientists are now trying to locate and identify the function of all the genes along the sequence.

1.1.4 Miscellaneous facts

Each chromosome is one continuous strand of DNA. The human genome is 46 chromosomes – 23 pairs. The best current estimate is that this is approximately 23,000 genes. The complete genome is approximately 3 billion base pairs long.

Higher cells also have a mitrochondrial chromosome that is maternally inherited.

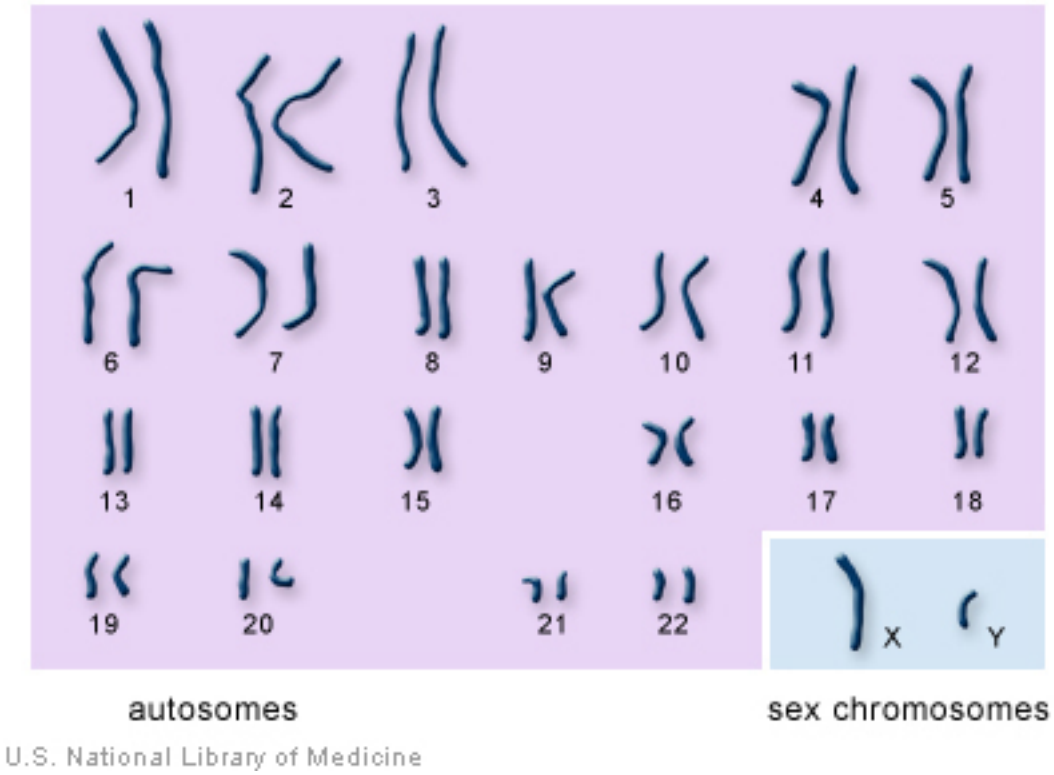


Figure 1.4: The human genome.

1.2 Population Genetics

Hidden within the genetic code is compelling evidence of the shared ancestry of all living things. It is a record of all the mutations that have occurred and become fixed within the genome. The aim of population genetics is to study genetic variation and evolution to make conclusions about entire populations and their history. This could be about the growth of the cells in a tumor, the history of a small group of individuals such as a family, or a larger population, or a whole species, or a collection of species. We are still only at the beginning of understanding and being able to read the history that is embedded within the genome.

Examples of the types of question we might want to answer are

- When did humans diverge from chimps? See Figure 1.5
- Did a population experience a drastic bottleneck in

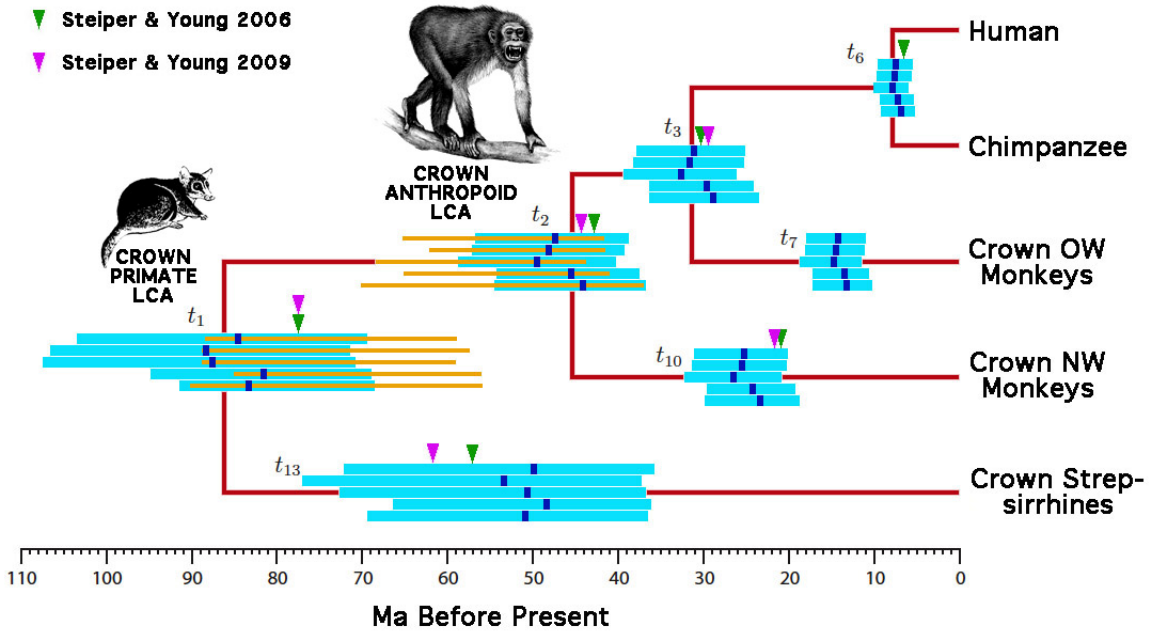


Figure 1.5: Primate evolution.

the past?

- Are two distinguishable populations mixing genetically? How much?
- Why did sex evolve?
- What impact does inbreeding have on a population? For example, how much mutational damage has consanguineous marriages in human populations caused?

Population genetics is a subject that is unimaginable without sophisticated mathematics and statistics. Some of the greatest names in mathematics and statistics have worked in the field, including Fisher, Galton, Pearson, Hardy, ...

1.3 Genetics Glossary

Genetics is the study of *genes*.

- Alleles: Alternative forms of a genetic locus; a single allele for each locus is inherited separately from each parent (e.g., at a locus for eye color the allele might result in blue or brown eyes).

- Amino acid: Any of a class of 20 molecules that are combined to form proteins in living things. The sequence of amino acids in a protein and hence protein function are determined by the genetic code.

- Base pair (bp): Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

- Base sequence: The order of nucleotide bases in a DNA molecule.

- Centimorgan (cM): A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million base pairs.

- Chromosomes: The self- replicating genetic structures of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.

- Crossing over: The breaking during meiosis of one maternal and one paternal chromosome, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of alleles between chromosomes. Compare recombination.

- DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double- stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cyto-

sine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner.

- DNA sequence: The relative order of base pairs, whether in a fragment of DNA, a gene, a chromosome, or an entire genome. See base sequence analysis. Double helix: The shape that two linear strands of DNA assume when bonded together.
- Double helix: The shape that two linear strands of DNA assume when bonded together.
- Eukaryote: Cell or organism with membrane- bound, structurally discrete nucleus and other well- developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue- green algae. Compare prokaryote. See chromosomes.
- Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule).
- Genetics: The study of the patterns of inheritance of specific traits.
- Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.
- Haploid: A single set of chromosomes (half the full set of genetic material), present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare diploid.
- Heterozygosity: The presence of different alleles at one or more loci on homologous chromosomes.
- Kilobase (kb): Unit of length for DNA fragments equal to 1000 nucleotides.
- Linkage: The proximity of two or more markers (e.g.,

genes, RFLP markers) on a chromosome; the closer together the markers are, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.

- Linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans (cM).
- Locus (pl. loci): The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. The use of locus is sometimes restricted to mean regions of DNA that are expressed. See gene expression.
- Marker: An identifiable physical location on a chromosome (e.g., restriction enzyme cutting site, gene) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined. See RFLP, restriction fragment length polymorphism.
- Megabase (Mb): Unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM.
- Meiosis: The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes.
- Mitosis: The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell.
- Mutation: Any heritable change in DNA sequence. Compare polymorphism.
- Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine,

or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. See DNA, base pair, RNA.

- Nucleus: The cellular organelle in eukaryotes that contains the genetic material.
- Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest- resolution physical map is the banding patterns on the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.
- Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic linkage analysis. Compare mutation.
- Prokaryote: Cell or organism lacking a membrane-bound, structurally discrete nucleus and other subcellular compartments. Bacteria are prokaryotes. Compare eukaryote. See chromosomes.
- Protein: A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has unique functions. Examples are hormones, enzymes, and antibodies.
- Recombination: The process by which progeny derive a combination of genes different from that of either parent. In higher organisms, this can occur by crossing over.
- Sequence: See base sequence.

- Sequencing: Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.
- Sex chromosomes: The X and Y chromosomes in human beings that determine the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a karyotype. Compare autosome.
- Single- gene disorder: Hereditary disorder caused by a mutant allele of a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare polygenic disorders.

Chapter 2

Hardy-Weinberg Equilibrium

Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

Population genetics is the study of allele frequency distributions and changes under the influence of four evolutionary processes:

1. Natural selection
    - process whereby heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations. Natural selection acts on the phenotype, and can lead to the development of new species.
2. Genetic drift
    - changes in relative frequency of an allele due to random sampling and chance (not driven by environmental or adaptive processes). Changes can be beneficial, neutral or detrimental. The effect is larger in smaller populations.
3. Mutation
    - changes in DNA caused by radiation, viruses, replication errors, etc. Changes can be bene-

ficial, neutral or detrimental. Usual mutation rate in mammals is extremely low, about 1 in 10^7 bases. Some viruses benefit from a high mutation rate, allowing them to evade immune systems.

4. Gene flow

    - exchanges of genes between populations, for example by migration and breeding. Hindered by mountains, oceans, deserts, Great Wall of China, etc. Acts strongly against speciation.

Population genetics has a long and heated history of debates between scientists who argue for the primacy of the different causes of frequency changes. Before we can understand the mechanisms that cause a population to evolve, we must consider what conditions are required for a population not to evolve.

2.0.1 Types of genetic variation

Lets distinguish between three types of variation within a population

1. The number of alleles at a locus
2. The frequency of alleles at a locus
3. The frequency of genotypes at a locus

The first type of variation can only change by mutation or immigration. The second and third types of variation can change solely by genetic drift, as well as being driven by external forces.

Why do we need both of variation type 2 and 3 above?

We can always calculate allele frequencies from genotype frequencies, but we can't do the reverse unless ....

2.1 The Hardy-Weinberg Law

The Hardy-Weinberg law is a zero-force law (like Newton's first law of dynamics). It says that in the absence of any driving force on the population, allele frequencies remain constant. This is an idealised law, which depends on many assumptions which are never met in reality, however it is important as deviations from Hardy-Weinberg (HW) equilibrium can suggest what forces are acting on the population.

Lets consider a single locus where there are just two alleles segregating in a diploid population. The Hardy-Weinberg assumptions are

1. No difference in gentype propotions between the sexes
2. Synchronous reproduction at discrete points in time (discrete generations)
3. Infinite population size (so that frequencies can be replaced by expectations)
4. No mutation
5. No migration (no immigration and balanced emigration)
6. No selection (your genotype does not influence the probability that you reproduce)
7. Random mating (wrt their genotype at this particular locus)

Suppose the two alleles are denoted A and a, and that the genotype frequences are

Genotype	AA	Aa	aa
Frequency	X	$2Y$	Z

Since we have random matings,

- the frequency of matings of the type $AA \times AA$ is X^2, and the outcome is always zygotes with genotype AA.
- the frequency of $AA \times Aa$ matings is $4XY$ (why?), and the rules of Mendelian inheritance tell us that this produces AA zygotes with probability 0.5, and Aa zygotes with probability 0.5.
- the frequency of $Aa \times Aa$ matings is $4Y^2$, and this produces AA and aa with probability 0.25 each, and Aa with probability 0.5.
- ...
- ...
- ...

Lets now consider the frequency genotypic frequencies in the next generation, which we denote as X' for the frequency of AA, $2Y'$ for Aa and Z' for aa.

$$\begin{aligned} X' &= X^2 + \frac{1}{2}(4XY) + \frac{1}{4}(4Y^2) \\ &= (X+Y)^2 \end{aligned} \tag{2.1}$$

Similarly,

$$\begin{aligned} 2Y' &= \frac{1}{2}(4XY) + \frac{1}{2}(4Y^2) + 2XZ + \frac{1}{2}(4YZ) \\ &= 2(X+Y)(Y+Z) \end{aligned} \tag{2.2}$$

$$\begin{aligned} Z' &= Z^2 + \frac{1}{2}(4YZ) + \frac{1}{4}(4Y^2) \\ &= (Y+Z)^2 \end{aligned} \tag{2.3}$$

If we consider the next generation, we can find frequencies X'', Y'' and Z'' by replacing X by X'. Y by Y' etc in the equations above. This gives

$$\begin{aligned} X'' &= (X' + Y')^2 \text{ by Equation (2.1)} \\ &= (X + Y)^2 \text{ by Equation (2.2) – why?} \\ &= X' \end{aligned}$$

Similarly,

$$\begin{aligned} Y'' &= \\ &= \\ &= Y' \end{aligned}$$

and

$$\begin{aligned} Z'' &= \\ &= \\ &= Z' \end{aligned}$$

Thus the genotypic frequencies established by the second generation are maintained in the third generation and consequently in all subsequent generations.

Frequencies having this property can be characterised as those satisfying the relation

$$(Y')^2 = X'Z'$$

Clearly, if this also holds for the first generation

$$Y^2 = XZ, \tag{2.4}$$

then the frequencies will be the same in all generations. Populations for which (2.4) is true are said to be in Hardy-Weinberg equilibrium (HWE).

Why is this law important?

- Firstly, HWE says that if no external forces act, there is no tendency for variation to disappear. This immediately showed that the early criticism of Darwinism, namely that variation decreases rapidly under blending theory, does not apply with Mendelian inheritance. In fact, HW shows that under a Mendelian

system of inheritance variation tends to be maintained (of course, selection will tend to remove variation). Note that it is the 'quantal' nature of the gene that leads to the stability behaviour described by HW. It is has been argued that if there is intelligent life evolved by natural selection elsewhere in the universe, the heredity mechanism involved must be a quantal one (maybe Mendelian), since otherwise it is not clear how the required variation for natural selection could be maintained. It may be a coincidence, but Quantum theory in physics was propsed in the same year (1900) that Mendelian inheritance was rediscovered.

- Secondly, HWE allows us (for populations which may reasonably be assumed to undergo random mating) to consider just the variation of alleles at a locus, rather than having to consider the genotype frequencies.

  Since $X + 2Y + Z = 1$ we know only two of the frequencies are independent. If Equation (2.4) also holds, only one frequency is independent. Let x be the frequency of the allele A. Then we can summarise the Hardy-Weinbery law as

  Theorem 1 *(Hardy-Weinberg 1908) Under the stated assumptions, a population having genotypic frequencies $X, 2Y$ and Z corresponding to genotypes AA, Aa, aa respectively, achieves, after one generation of random mating, stable genotypic frequences x^2, $2x(1-x)$, $(1-x)^2$, where $x = X + Y$ and $1 - x = Y + Z$. If the initial frequencies are already of this form, then these frequencies are stable for all generations.*

  Hence, the mathematical behaviour of the population can be examined in terms of the single frequency x, rather than the triplet (X, Y, Z).

- Finally, if genotypes are not in Hardy-Weinberg proportions, one or more of the 8 assumptions **must** be false! Departures from HWE are one way in which

we can detect evolutionary forces within populations and estimate their magnitude.

Generalisation to multiple alleles

Suppose there are k different alleles, $A_1, \ldots, A_k$ with population frequencies $p_1, \ldots, p_k$. Then, upon random union, the diploid frequencies are

$$P_{ii} = p_i^2 \text{ for } i = 1, 2, \ldots, k$$
$$P_{ij} = 2p_i p_j \text{ for } i \neq j$$

See questions 1, 2 and 4 on the example sheet.

2.2 Estimating Allele Frequencies

In order to make statements about the genetic structure of populations, we need to be able to count the number of genotypes and alleles. However, for simple Mendelian traits, it is not possible to directly know someone's genotype unless they possess the recessive trait. If one allele is dominant over the other, then it means that there is a

strong phenotypic effect in heterzygotes that conceals the presence of the weaker allele.

Consider the famous example of industrial melanism in moths. As pollution increased in Great Britain, several species of moths evolved black camouflage when resting, so that they would merge into the coal soot produced during the industrial revolution. In most instances, the melanic colour pattern has been found to be due to a single dominant allele. In a 1956 study in a region of Birmingham, it was found that 87% of the moth species *Biston betularia* had the melanic colour pattern. In otherwords, 13% of moths were found to be homozygous recessives.

How do we determine the frequency of the dominant allele in the population? We don't directly observe the number of heterozygous moths, so we can't simply determine the frequency of the recessive allele. However, if we are prepared to assume random mating in the population, then we can use the Hardy-Weinberg equilibirum proportions to determine the frequencies.

Let R denote the number of homozygous recessive carriers, and let q be the frequency of the recessive allele. The HW proportions are

AA	Aa	aa
$(1-q)^2$	$2q(1-q)$	q^2

If we sampled N moths in total, then R will have a binomial

distribution

$$\mathbb{P}(R|q) = \binom{N}{R}(q^2)^R(1-q^2)^{N-R}$$

We will estimate q by its maximum likelihood estimate, found by taking the log of the above likelihood, differentiating, and setting the derivative to zero.

This gives

$$\hat{q} = \sqrt{\frac{R}{N}}$$

So for the moths, we observed $R/N = 0.13$, which gives an estimate of $\hat{q} = \sqrt{0.13} = 0.36$ for the recessive allele, and 0.64 for the dominant allele. The estimate of frequencies of the dominant homozygotes, heterozygotes, and recessive homozygotes are 0.41, 0.46 and 0.13 respectively.

Which Hardy-Weinberg assumptions might be violated here?

2.2.1 Fisher's approximate variance formula

How confident are we in these estimates? One way to answer this is to compute variances of our estimates. One way to determine the variance of a frequency estimator is to use Fisher's approximate variance formula.

Suppose we are given count data $n_1, n_2, \ldots, n_k$, with $n = n_1 + \ldots + n_k$, which come from a multinomial distribution with parameters $q_1, \ldots q_k$.

$$\mathbb{P}(n_1, \ldots, n_k) = \frac{n!}{\prod n_i!} \prod q_i^{n_i}$$

Fisher's approximate variance function then says that if $T = T(n_1, \ldots, n_k)$ is a function of the counts n_i, then

$$\mathbb{V}\mathrm{ar}(T) \approx n \sum_i \left(\frac{\partial T}{\partial n_i}\right)^2 q_i - n \left(\frac{\partial T}{\partial n}\right)^2$$

where q_i is the true value of the i^{th} parameter. In practice, we have to replace q_i by an estimator of it. The derivatives $\frac{\partial T}{\partial n_i}$ should be evaluated at their expected values, namely with $n_i = q_i n$.

The second term is only needed when T explicitly involves the sample size n. The above approximation works when either T is a ratio of functions of the same order in the counts, or when the counts n_i in T only appear divided by the total sample size n.

Derivation

Fisher's variance formula is derived using the **Delta Method** (first seen in G12SMM). Suppose we have an estimator $\mathbf{B}$ of the k-dimensional parameter $\boldsymbol{\beta}$. Suppose further that $\mathbf{B}$ is a consistent estimator which converges in probability to its true value $\boldsymbol{\beta}$ as the sample size $n \to \infty$. Suppose we want to estimate the variance of a function T of the estimator $\mathbf{B}$. Keeping only the first two terms of the Taylor series, and using vector notation for the gradient, we can estimate $T(\mathbf{B})$ as

$$T(\boldsymbol{\beta} + (\mathbf{B} - \boldsymbol{\beta})) = T(\mathbf{B}) \approx T(\boldsymbol{\beta}) + \nabla T(\boldsymbol{\beta})^t (\mathbf{B} - \boldsymbol{\beta})$$

where $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, and

$$\nabla T(\boldsymbol{\beta})^t = \left(\frac{\partial T}{\partial \boldsymbol{\beta}_1}, \ldots, \frac{\partial T}{\partial \boldsymbol{\beta}_k}\right)$$

This implies that the variance of $T(\mathbf{B})$ is approximately

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(T(\mathbf{B})\right) &\approx \mathbb{V}\mathrm{ar}\left(T(\boldsymbol{\beta}) + \nabla T(\boldsymbol{\beta})^t(\mathbf{B}-\boldsymbol{\beta})\right)\\ &= \mathbb{V}\mathrm{ar}\left(T(\boldsymbol{\beta}) + \nabla T(\boldsymbol{\beta})^t\mathbf{B} - \nabla T(\boldsymbol{\beta})^t\boldsymbol{\beta}\right)\\ &= \mathbb{V}\mathrm{ar}\left(\nabla T(\boldsymbol{\beta})^t\mathbf{B}\right)\\ &= \nabla T(\boldsymbol{\beta})^t\mathbb{V}\mathrm{ar}(\mathbf{B})\nabla T(\boldsymbol{\beta})\end{aligned}$$

where $\mathbb{V}\mathrm{ar}(\mathbf{B})$ is a $k\times k$ matrix with ij^{th} entry $\mathbb{C}\mathrm{ov}(B_i, B_j)$.
We can write this product as a sum, to find

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(T(\mathbf{B})\right) &\approx \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial T}{\partial\beta_i}\frac{\partial T}{\partial\beta_j}\mathbb{C}\mathrm{ov}(B_i,B_j)\\ &= \sum_{i=1}^{k}\left(\frac{\partial T}{\partial\beta_i}\right)^2\mathbb{V}\mathrm{ar}(B_i) + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\frac{\partial T}{\partial\beta_i}\frac{\partial T}{\partial\beta_j}\mathbb{C}\mathrm{ov}(B_i,B_j)\end{aligned}$$

If we specialise the above formula to the case where $T = T(n_1,\ldots,n_k)$, and assume that $(n_1,\ldots,n_k)\sim\mathrm{Multinomial}(n;q_1,\ldots,q_k)$. Let $\beta_i = q_i$ and note that $B_i = n_i/n$ is an estimator of q_i, with $n_i/n\to q_i$ as $n\to\infty$. We can then rewrite the delta formula as

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(T(\mathbf{B})\right) &\approx \sum_{i=1}^{k}\left(\frac{\partial T}{\partial n_i}\frac{\partial n_i}{\partial q_i}\right)^2\mathbb{V}\mathrm{ar}\left(\frac{n_i}{n}\right)\\ &\quad + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\left(\frac{\partial T}{\partial n_i}\frac{\partial n_i}{\partial q_i}\right)\left(\frac{\partial T}{\partial n_j}\frac{\partial n_j}{\partial q_j}\right)\mathbb{C}\mathrm{ov}\left(\frac{n_i}{n},\frac{n_j}{n}\right)\\ &= \sum_{i=1}^{k}\left(\frac{\partial T}{\partial n_i}\right)^2\mathbb{V}\mathrm{ar}\left(n_i\right) + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\frac{\partial T}{\partial n_i}\frac{\partial T}{\partial n_j}\mathbb{C}\mathrm{ov}\left(n_i,n_j\right)\end{aligned}$$

where we evaluate the derivatives with $n_i = \mathbb{E}(n_i) = nq_i$.
This then gives $\frac{\partial n_i}{\partial q_i} = n$ and so cancels with the $1/n^2$ from $\mathbb{C}\mathrm{ov}(n_i/n, n_j/n) = 1/n^2\mathbb{C}\mathrm{ov}(n_i,n_j)$

Variances and covariances for the multinomial distribution can be shown to be

$$\begin{aligned}\mathbb{V}\mathrm{ar}(n_i) &= nq_i(1-q_i)\\ \mathbb{C}\mathrm{ov}(n_i, n_j) &= -nq_iq_j\end{aligned}$$

This gives

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(T(\mathbf{B})\right) &\approx n\sum_{i=1}^{k}\left(\frac{\partial T}{\partial n_i}\right)^2 q_i(1-q_i) - n\sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\frac{\partial T}{\partial n_i}\frac{\partial T}{\partial n_j}q_iq_j\\ &= n\sum_{i=1}^{k}\left(\frac{\partial T}{\partial n_i}\right)^2 q_i - n\left(\sum_{i=1}^{k}\frac{\partial T}{\partial n_i}q_i\right)^2\end{aligned}$$

Finally, noting that

$$\begin{aligned}\frac{\partial T}{\partial n} &= \sum_{i=1}^{k}\frac{\partial T}{\partial n_i}\frac{\partial n_i}{\partial n}\\ &= \sum_{i=1}^{k}\frac{\partial T}{\partial n_i}q_i\end{aligned}$$

gives the required result.

Moths example

In this case, our estimator is

$$T = \hat{q} = \sqrt{\frac{R}{N}}$$

where R is our count. Fisher's variance formula gives

$$\mathbb{V}\mathrm{ar}(T) \approx N\left(\frac{\partial T}{\partial R}\right)^2 q^2 - N\left(\frac{\partial T}{\partial N}\right)^2$$

We can see that

$$\left.\frac{\partial T}{\partial R}\right|_{R=q^2N} = \frac{1}{2Nq} \tag{2.5}$$

$$\left.\frac{\partial T}{\partial N}\right|_{R=q^2N} = -\frac{q}{2N} \tag{2.6}$$

which gives

$$\mathbb{V}\mathrm{ar}(T) = \mathbb{V}\mathrm{ar}(\hat{q}) \tag{2.7}$$

$$\approx N\left[\left(\frac{1}{2Nq}\right)^2 q^2 - \left(\frac{q}{2N}\right)^2\right] \tag{2.8}$$

$$= \frac{1}{4N} - \frac{q^2}{4N} \tag{2.9}$$

We don't know q of course, but we can substitute in the estimate $\hat{q}$ to get

$$\mathbb{V}\mathrm{ar}(\hat{q}) = \frac{1}{4N}\left(1 - \frac{R}{N}\right)$$

If there were 100 moths in our study, this gives

$$\mathbb{V}\mathrm{ar}(\hat{q}) = 87/(4 * 100^2) = 0.0022$$

So our estimate of the recessive allele frequency is

$$0.36 \pm 0.047$$

where $se(\hat{q}) = 0.046$.

2.3 The EM algorithm

The EM algorithm is an iterative procedure for maximum likelihood estimation. It is primarily useful when there is missing data. By thinking about what kind of additional data it would be useful to have, it is possible to make the calculation simpler.

Suppose we write the likelihood function as

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{x}$ denotes *incomplete* data.

Suppose we can 'complete' $\boldsymbol{x}$ with $\boldsymbol{y}$, so that $(\boldsymbol{x}, \boldsymbol{y})$ is the complete data

We then consider the likelihood based upon $(\boldsymbol{x}, \boldsymbol{y})$, $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$.
To estimate the maximum likelihood estimate $\hat{\theta} = \arg\max L(\theta; x)$ we do the following:

We start at $\boldsymbol{\theta}^{(0)}$; then generate a sequence of iterates, $\boldsymbol{\theta}^{(m)}$. Each iteration consists of two steps.

$\boldsymbol{E}$-step (E stands for **expectation**) Calculate

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= \mathbb{E}_Y[\log L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\theta}^{(m)}] \\ &= \int_Y \log(L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})) f(\boldsymbol{y}|\theta^{(m)}, \boldsymbol{x}) \mathrm{d}\boldsymbol{y} \end{aligned}$$

where $f(y|\theta, x)$ is the pdf of y given θ and x.

$\boldsymbol{M}$-step (M stands for **maximisation**) Find the value of $\boldsymbol{\theta}$, called $\boldsymbol{\theta}^{(m+1)}$, which maximises $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ by solving

$$\frac{\partial Q}{\partial \theta_j}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = 0, \quad j = 1, \ldots, p.$$

The algorithm works by augmenting X with Y so that the likelihood $L(\theta|X, Y)$ is easier to maximise. We then

1. Estimate Y from a current estimate of θ given $X = x$
2. Maximise $L(\theta|x, y)$ with respect to θ

and iterate to convergence.

The EM algorithm produces a sequence of values that converge to a stationary value.

- If the likelihood is unimodal it will converge to the MLE.
- In general, there is no guarantee that it will give the MLE, only that it will reach a maximum (which might be a local maximum).

For proofs of convergence see G14CST.

2.3.1 Recessive allele frequency estimation revisited

In order to consider a simple version of the EM algorithm (before the more complex example to follow), consider the two-allele considered before.

AA	Aa	aa
$(1-q)^2$	$2q(1-q)$	q^2

We observe n_{aa} (the number of recessive homozygotes) and n_d (the number of dominant homozygotes plus the heterozygotes), but we don't observe N_{AA} or N_{Aa}, only their sum n_d. So we can consider N_{AA} and N_{Aa} as missing.

The likelihood of the complete data is proportional to

$$((1-q)^2)^{N_{AA}}(2q(1-q))^{N_{Aa}}(q^2)^{n_{aa}}$$

Here we use upper case letters to emphasise that the variables N_{AA} and N_{Aa} are unknown, but that n_{aa} is known. The log-likelihood is thus proportional to

$$2N_{AA}\log(1-q) + N_{Aa}\log(q(1-q)) + 2n_{aa}\log(q)$$

We must take the expectation of this with respect to the distribution of N_{AA} and N_{Aa} given $q^{(m)}$ and n_{aa} and n_d. It is easy to see that

$$\begin{aligned}
\mathbb{E}(N_{AA}|q(m), n_{aa}, n_d) &= n_d \frac{(1-q(m))^2}{(1-q(m))^2 + 2q(m)(1-q(m))} \\
&= n_d \frac{1-q(m)}{1+q(m)} \\
&= n_{AA}(m) \text{ say} \\
\mathbb{E}(N_{Aa}|q(m), n_{aa}, n_d) &= n_d \frac{2q(m)(1-q(m))}{(1-q(m))^2 + 2q(m)(1-q(m))} \\
&= n_d \frac{2q(m)}{1+q(m)} \\
&= n_{Aa}(m) \text{ say}
\end{aligned}$$

So when we take the expectation of the log-likelihood above we get

$$2n_{AA}(m)\log(1-q) + n_{Aa}(m)\log(q(1-q)) + 2n_{aa}\log(q)$$

which can be maximised by differentiating, setting equal to zero and solving for q. This gives

$$\begin{aligned} q(m+1) &= \frac{2n_{aa} + n_{Aa}(m)}{2n_{AA}(m) + 2n_{Aa}(m) + 2n_{aa}} \\ &= \frac{2n_{aa} + n_{Aa}(m)}{2n} \end{aligned}$$

To summarise, the EM algorithm for finding the recessive allele frequency in the two allele problem iterates through the following two equations until convergence

$$n_{Aa}(m) = n_d \frac{2q(m)}{1 + q(m)} \tag{2.10}$$

$$q(m+1) = \frac{2n_{aa} + n_{Aa}(m)}{2n} \tag{2.11}$$

Lets apply this to the data above to check we get the same answer. Recall that $n_d = 87$ and $n_{aa} = 13$. Lets initialise the algorithm with $q(0) = 0.5$ (you should try other start values)

Iteration m	$n_{Aa}(m)$	$q(m)$
0	–	0.5
1	58.00	0.42
2	51.4648	0.3873
3	48.5787	0.3729
4	47.2604	0.3663
5	46.6489	0.3632
6	46.3633	0.3618
7	46.2295	0.3611
8	46.1667	0.3608
9	46.1372	0.3607
10	46.1233	0.3606

which agrees with our previous answer.

The reason for studying the EM algorithm, rather than just directly maximising the likelihood, is that the EM algorithm can be used in more complex situations where direct maximisation is not possible.

2.3.2 More complex allele estimation problems

Consider the ABO blood group system in humans. There are 4 distinct phenotypes distinguished and 6 different genotypes.

Phenotype	A	AB	B	O
Genotype(s)	aa ao	ab	bb bo	oo
No. in sample	n_A	n_{AB}	n_B	n_O

Suppose we want to know the allele frequencies. There are 3 alleles in total: a, b and o. a and b are codominant, but both dominant over o. Let p_a be the frequency of allele a in the population, similarly for p_b and p_o. We want to estimate these proportions.

Lets consider the missing data in this case to be N_{aa} and N_{bb}. The complete data is then $(n_A, N_{aa}, n_{AB}, n_B, N_{bb}, n_O)$, and the likelihood of the complete data is proportional to

$$(p_a^2)^{N_{aa}}(2p_ap_o)^{n_A-N_{aa}}(2p_ap_b)^{n_{AB}}(p_b^2)^{N_{bb}}(2p_bp_o)^{n_B-N_{bb}}(p_o^2)^{n_O}$$

which makes the log-likelihood

$$2N_{aa}\log(p_a)+(n_A-N_{aa})\log(2p_ap_o)+n_{AB}\log(2p_ap_b)+2N_{bb}\log(p_b)+(n_B-N_{bb})\log(2p_bp_o)+2n_O\log(p_o)$$

The EM algorithm requires us to take expectations with respect to the distribution of N_{aa} and N_{bb} given n_A, n_B, n_{AB}, n_O and $p_a(m), p_b(m), p_o(m)$. As N_{aa} and N_{bb} only occur as simple linear combinations, we only require $\mathbb{E}N_{aa}$ and $\mathbb{E}N_{bb}$ given n_A, n_B, n_{AB}, n_O and $p_a(m), p_b(m), p_o(m)$. The expectations can easily be seen to be

$$\begin{aligned}\mathbb{E}N_{aa} &= n_A\frac{p_a(m)^2}{p_a(m)^2+2p_a(m)p_o(m)}\\ &= n_{aa}(m) \text{ say}\\ \mathbb{E}N_{bb} &= n_B\frac{p_b(m)^2}{p_b(m)^2+2p_b(m)p_o(m)}\\ &= n_{bb}(m) \text{ say}\end{aligned}$$

Let $n_{ao}(m) = n_A - n_{aa}(m)$ and $n_{bo}(m) = n_B - n_{bb}(m)$.
Substituting these into the loglikelihood gives

$$\begin{aligned}2n_{aa}(m)\log(p_a) + n_{a0}(m)\log(2p_ap_o) + n_{AB}\log(2p_ap_b)\\ + 2n_{bb}(m)\log(p_b) + n_{bo}(m)\log(2p_bp_o) + 2n_O\log(p_o)\end{aligned}$$

We want to maximise this likelihood subject to the constraint that

$$p_a + p_b + p_0 = 1.$$

Using Lagrange multipliers, we can do this by maximixing

$$\begin{aligned}2n_{aa}(m)\log(p_a) + n_{a0}(m)\log(2p_ap_o) + n_{AB}\log(2p_ap_b)\\ + 2n_{bb}(m)\log(p_b) + n_{bo}(m)\log(2p_bp_o) + 2n_O\log(p_o)\\ + \lambda(1 - p_a - p_b - p_o)\end{aligned}$$

Differentiating with respect to p_a, p_b, p_0 and λ and setting the derivatives to 0 gives

$$\begin{aligned}\lambda &= \frac{1}{p_a}(2n_{aa}(m) + n_{ao}(m) + n_{AB})\\ \lambda &= \frac{1}{p_b}(n_{bo}(m) + n_{AB} + 2n_{bb})\\ \lambda &= \frac{1}{p_o}(n_{ao}(m) + n_{bo}(m) + 2n_O)\\ 1 &= p_a + p_b + p_o\end{aligned}$$

Rearranging and adding the three equations gives

$$\begin{aligned}\lambda(p_a + p_b + p_o) &= \lambda\\ &= 2n_{aa}(m) + 2n_{AB} + 2n_{ao}(m) + 2n_{bb}(m) + 2n_{bo}(m) + 2n_O\\ &= 2n\end{aligned}$$

where n is the total number of subjects in the dataset.
Thus, we find

$$\begin{aligned}p_a(m+1) &= \frac{2n_{aa}(m) + n_{ao}(m) + n_{AB}}{2n}\\ p_b(m+1) &= \frac{2n_{bb}(m) + n_{bo}(m) + n_{AB}}{2n}\\ p_o(m+1) &= \frac{2n_O + n_{ao}(m) + n_{bo}(m)}{2n}\end{aligned}$$

To summarize, we can estimate the allele frequencies for the ABO blood groups by first picking an initial value of p_a, p_b, p_o, then iterating through the following equations:

$$n_{aa}(m) = n_A \frac{p_a(m)^2}{p_a(m)^2 + 2p_a(m)p_o(m)} \tag{2.12}$$

$$n_{ao}(m) = n_A - n_{aa}(m) \tag{2.13}$$

$$n_{bb}(m) = n_B \frac{p_b(m)^2}{p_b(m)^2 + 2p_b(m)p_o(m)} \tag{2.14}$$

$$n_{bo}(m) = n_B - n_{bb}(m) \tag{2.15}$$

$$p_a(m+1) = \frac{2n_{aa}(m) + n_{ao}(m) + n_{AB}}{2n} \tag{2.16}$$

$$p_b(m+1) = \frac{n_{bo}(m) + n_{AB} + 2n_{bb}(m)}{2n} \tag{2.17}$$

$$p_o(m+1) = \frac{n_{ao}(m) + n_{bo}(m) + 2n_O}{2n} \tag{2.18}$$

Lets try this on real data. On one test of 6313 Caucasians in Iowa City, the numbers of individuals with blood types A, B, O and AB were found to be

Phenotype	A	AB	B	O
Observed frequency	2625	226	570	2892

If we initialise the algorithm with $p_a = 1, p_b = 0, p_o = 0$ (you should try other starting points), we find we soon reach convergence.

Iteration m	$p_a(m)$	$p_b(m)$	$p_o(m)$
0	1	0	0
1	0.4337	0.1082	0.4581
2	0.2926	0.0678	0.6396
3	0.2645	0.0653	0.6702
4	0.2601	0.0651	0.6748
5	0.2594	0.0651	0.6755
6	0.2593	0.0651	0.6756
7	0.2593	0.0651	0.6756
8	0.2593	0.0651	0.6756

The R code for this operation is on the course website.

2.4 Testing for HWE

We can think of the Hardy-Weinberg proportions as being a hypothesis about the structure of the population. We then want to test whatever observations we observe, to see if there is evidence of departure from HWE.

Recall from G12SMM or G14FOS, the likelihood ratio test. Suppose we have a statistical model of the data that depends on unknown parameter $\boldsymbol{\theta} \in \Theta$, described by likelihood function $\mathcal{L}(\boldsymbol{\theta})$. Suppose we want to test the following two hypotheses about $\boldsymbol{\theta}$:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ vs } H_1 : \boldsymbol{\theta} \notin \Theta_0$$

where $\Theta_0 \subset \Theta$. The likelihood-ratio statistic is

$$\Lambda = 2 \log \left(\frac{\sup_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_0} \mathcal{L}(\boldsymbol{\theta})} \right) = 2 \log \left(\frac{\mathcal{L}(\hat{\boldsymbol{\theta}})}{\mathcal{L}(\hat{\boldsymbol{\theta}}_0)} \right)$$

where $\hat{\boldsymbol{\theta}}$ is the MLE and $\hat{\boldsymbol{\theta}}_0$ is the MLE when $\boldsymbol{\theta}$ is restricted to lie in Θ_0.

Wilks' theorem then says that under certain regularity conditions and for n large, that under H_0

$$2 \log \Lambda \sim \chi^2_{p_1 - p_0}$$

where $p_1 = \dim \Theta$ and $p_0 = \dim \Theta_0$.

For multinomial models, i.e., where we assume the data have a multinomial distribution with parameter $\boldsymbol{\theta}$, that is $(n_1, n_2, \ldots, n_k) \sim$ Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$, we can usually make a further approximation and use Pearson's chi-squared statistic:

$$2 \log \Lambda \approx \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the i^{th} observed count and E_i is the the expected count under H_0.

2.4.1 Blood group example

Consider again the observed frequencies of the blood group system:

Phenotype	A	AB	B	O
Genotype(s)	aa ao	ab	bb bo	oo
No. in sample	2625	226	570	2892

Lets test the following two hypotheses:

H_0 : the sample is in HWE vs H_1 : the sample is not in HWE

Saying the sample is in HWE is equivalent to saying we expect the following frequencies

Phenotype	A	AB	B	O
Genotype(s)	aa ao	ab	bb bo	oo
HW frequencies	$p_a^2 + 2p_a p_o$	$2p_a p_b$	$p_b^2 + 2p_b p_o$	p_o^2

The likelihood ratio test says we should substitute unknown frequencies $\boldsymbol{\theta} = (p_a, p_b, p_o)$ by their maximum likelihood estimates to estimate the expected phenotype frequencies under H_0. We found these in the previous section to be $\hat{\boldsymbol{\theta}}_0 = (0.2593, 0.0651, 0.6756)$. This gives

Phenotype	A	AB	B	O
Observed frequency	2625	226	570	2892
Expected frequency under H_0	2636.3	213.1	582.0	2881.4

These two sets of figures look like they closely match, but lets perform Pearson's chi-squared test to be sure. We find

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 1.11$$

To find the degrees of freedom, we need to consider the number of free parameters.

- For Θ_0, we constrain the four phenotype type frequencies to be in their HW form - but we still need to estimate p_a, p_b and p_o. This is only 2 free parameters, not 3, because we know they must add to 1. So $\dim \Theta_0 = 2$.

- For H_1, the only constraint is that the four phenotype frequencies add to 1. This gives us three free parameters, so $\dim \Theta = 3$.

So the degrees of freedom is $\dim \Theta - \dim \Theta_0 = 3 - 2 = 1$.

This gives that $T \sim \chi^2_1$. The upper 95th percentile of χ^2_1 is 3.8 and so there is not enough evidence to reject H_0 at the 5% level.

Notice that we haven't proven that the population is in HWE here (i.e. that H_0 is true), we have only shown that there is no evidence to suggest that it isn't true!

See Q5 and Q6 on the example sheet.

Chapter 3

Genetic Drift and Mutation

Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

The Hardy-Weinberg equilibrium frequencies were derived conditional upon a large number of assumptions. We described the HWL as being a bit like Newton's first law of motion, as it describes what happens to allele frequencies in the absence of any driving force. For the rest of the course, we will examine what happens when we violate some of the assumptions that were made before, such as

- Nonrandom mating
- Finite population size
- Non-zero selective forces
- Mutation of alleles

In this chapter we shall consider the effect of genetic drift due to finite population size, and mutation.

3.1 Genetic drift

The previous chapter on HWE assumed that we were dealing with an infinite population. This allowed us to deal only

with expected frequencies from random mating, and to ignore the stochastic effects of random sampling. For very large populations, this might be a good assumption, but for smaller populations, this is likely to lead to misleading results.

Genetic drift is the name given to the random changes in gene frequency that occur solely because of random sampling. To appreciate the effects of random sampling, lets consider the simplest model of drift, called the Wright-Fisher model.

3.1.1 Wright-Fisher model

Consider a random mating diploid population of N individuals, and lets focus on a single locus which has two naturally occuring alleles A and a.

In total, the population consists of $2N$ alleles, and we will treat this population as if it was a population of $2N$ haploid individuals. Lets assume that

1. Each individual in the population produces a large and identical number of gametes.
2. Each gamete is an identical copy of its parent
3. The next generation is constructed by picking $2N$ gametes at random from the large number originally produced.

These are the assumptions of the Wright-Fisher model. These assumptions are equivalent to assuming that we have a population of constant size $2N$. Each new generation, every individual picks its type by sampling with replacement from the types of the previous generation.

If we let X_t denote the number of A alleles in generation t, then we can see that X_t will be a Markov chain, taking one or other of the values $0, 1, \ldots, 2N$. Because we are assuming that the genes in generation $t+1$ are chosen with replacement from the genes of generation t, we can see that X_{t+1} will have a binomial distribution, with parameters $2N$

and $X_t/2N$. More explicitly, the transition probabilities of this chain can be seen to be

$$P_{rx} = \mathbb{P}(X_{t+1} = x | X_t = r) = \binom{2N}{x} \left(\frac{r}{2N}\right)^x \left(1 - \frac{r}{2N}\right)^{2N-x} \tag{3.1}$$

This Markov chain has two absorbing states (states from which we can never leave), and they are $X = 2N$ and $X = 0$. The first represents the fixation of allele A in the population, and the second represents the loss of A from the population.

One feature of genetic drift is that there is no systematic tendency for the frequency of alleles to move up or down. In other words, the nature of the random changes is neutral. To see this, note that

$$\mathbb{E}(X_{t+1}|X_t) = 2N\frac{X_t}{2N} = X_t. \tag{3.2}$$

It is also possible to show that the probability that any allele will eventually become fixed in the population is equal to its current frequency. This useful result is easy to prove but not done so here (see the section on martingale convergence in G14ASP for a proof). For an informal proof, note that eventually every gene in the population is descended from a unique gene in generation zero. The probability that suuch a gene is type A is simply the initial fraction of A genes, namely $X_0/2N$, and this must also be the fixation probability of A.

To help understand these dynamics, lets look at some simulations. Suppose $N = 20$ and that initially the frequency of A alleles is 20%. Then we can simulate from the Wright-Fisher model forward for 100 generations to generate sample trajectories through time illustrating genetic drift. Figure 3.1 shows 10 simulated trajectories.

Each line represents the change in allele frequency in a single population - no two trajectories are the same. We can see that in some trajectories the A allele is removed from the population (has frequency 0%), whereas in others it is fixed (frequency 100%). Note that once the frequency becomes fixed at 0 or 100% it cannot change in future

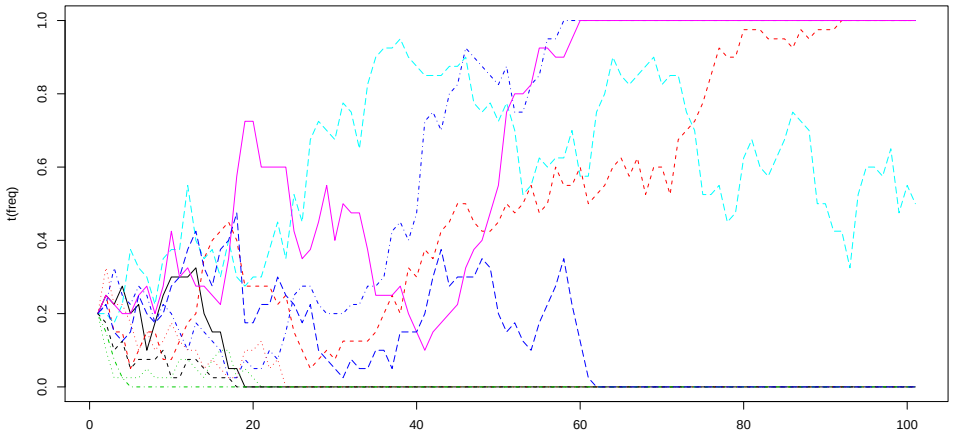


Figure 3.1: 10 simulated trajectories for a population of 20 diploid individuals. The initial gene frequency is 0.2 in all the trajectories.

generations. Only in 1 of the 10 trajectories is the frequency not fixed at either 0 or 1 after 100 generations.

The strength of genetic drift depends on the size of the population. We can see that

$$\mathbb{V}\mathrm{ar}(X_{t+1}|X_t) = 2N\frac{X_t}{2N}\left(1-\frac{X_t}{2N}\right) = \frac{X_t(2N-X_t)}{2N}$$

or if $p_t = X_t/(2N)$, then

$$\mathbb{V}\mathrm{ar}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N}. \tag{3.3}$$

So if N is large, we can see that the variance of the frequency in the next generation will be small, whereas if N is small, it will be large. In the limit as $N \to \infty$, we can see that the variance will be zero, indicating that there is no genetic drift. See Figure 3.2.

The effect of drift also depends on the current frequency of the allele. If p_t is close to 0 or 1, then the variance due to drift is small, whereas if $p = 1/2$, then the variance of drift is large.

Figure 3.1 illustrates another of the key effects of genetic drift. Namely, genetic drift acts to reduce variation from the population. Eventually, genetic drift will lead to the

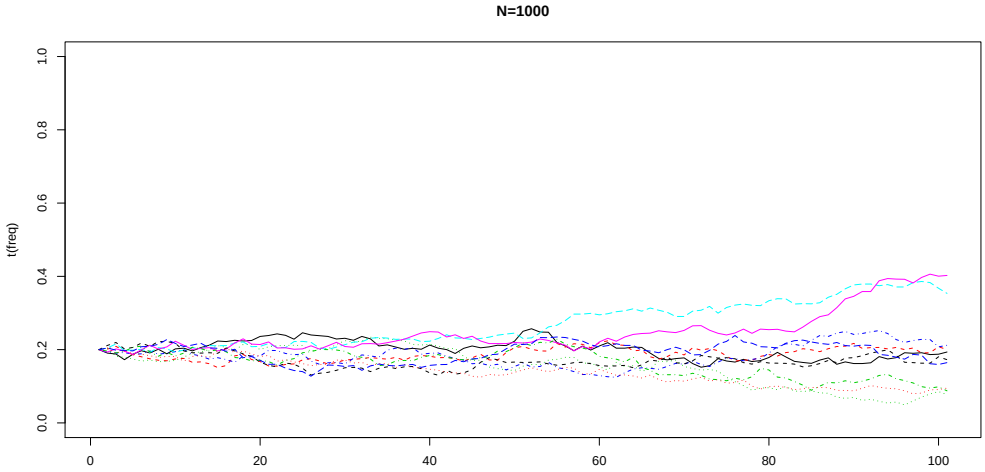


Figure 3.2: 10 simulated trajectories for a diploid population of size 1000. Compare this with Figure 3.1 to see the effect of N on genetic drift.

loss of all alleles in the population except one. Figure 3.3 shows the final frequency of allele A after a different number of generations in 10,000 simulated trajectories of this population. We will show this decay of heterozygosity mathematically in the next section.

Some comments:

- There are two main sources of randomness that lead to genetic drift. The first is Mendel's law of segregation and the second is the random number of offspring had by each individual.
- Although the Wright-Fisher model does not explicitly incorporate either of these two sources of randomness, it has been shown that more complex models that do model these terms behave in a very similar way to the Wright-Fisher model. Because the WF model is easy to understand, it is very commonly used.

Exercise:

What is the probability that a particular allele has a least one copy in the next generation? Show that as N increases this probability tends to $1 - 1/\mathrm{e}$.

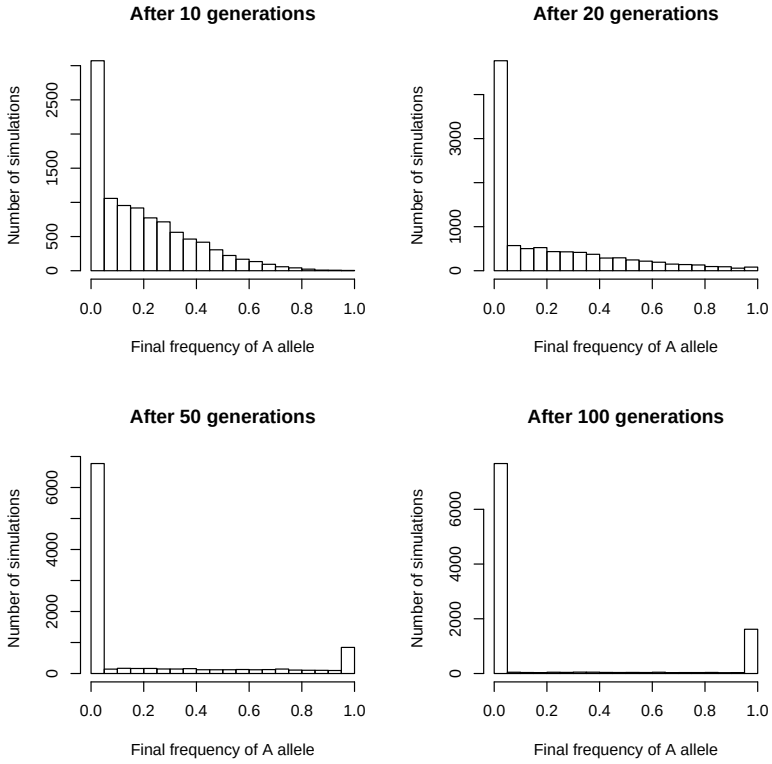


Figure 3.3: Final frequency of allele A after 10, 20 50 and 100 generations, for a population of 20 diploid individuals (10,000 simulated trajectories).

3.1.2 Decay of heterozygosity

Consider a diploid population of N hermaphroditic individuals. Let $\mathcal{G}$ be the probability that two alleles different by origin (drawn at random from the population without replacement) are identical by state. Then $\mathcal{G}$ measures the degree of genetic variation in the population;

$$\mathcal{G} = 0 \Rightarrow \text{ every allele is different by state}$$
$$\mathcal{G} = 1 \Rightarrow \text{ every allele is identical by state}$$

Then if we let $\mathcal{G}'$ be the value of $\mathcal{G}$ after one round of random mating, we can show that

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

If we then let $\mathcal{H} = 1 - \mathcal{G}$, so that $\mathcal{H}$ is the probability that two randomly drawn alleles are different by state ($\mathcal{H}$ is a measure of the heterozygosity of the population), then we can show that

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

and that

$$\Delta_N \mathcal{H} = -\frac{1}{2N}\mathcal{H}.$$

We can see that the probability that two alleles are different by state decreases at a rate of $1/(2N)$ for each generation, and thus will be extremely slow for large populations.

See Q2 from the example sheet.

We can see that most populations will end up homozygous, as genetic drift will tend to remove variation. However, the rate of change is very slow. For example, a population of 1 million individuals will require about 1.37×10^6 generations to reduce $\mathcal{H}$ by $1/2$. If the generation time of the species is 20 years (as in humans), then it would take 28 million years to halve the genetic variation.

3.1.3 Mean absortion time

One question we might want to ask is, what is the mean time to absorption of allele A (either at 0 or 100%) given initial frequency $X_0 = i$.

Propostion: Let $t(i)$ denote the mean time to absorption given $X_0 = i$. Then

$$t(i) \approx -4N(p \log p + (1-p)\log(1-p))$$

where $p = i/2N$.

Sketch of proof:

Let the initial frequency of allele A be $X_0 = i$. Consider the next generation X_1. Then

$$t(i) = \sum \mathbb{P}(X_1 = j|X_0 = i)t(j) + 1$$

as if $X_1 = j$, then the mean time to absorption is $t(j)$ and 1 generation has already passed. We can write

$$t(i) = \mathbb{E}(t(X_1)|X_0) + 1$$

Now lets convert to frequencies and write $p = X_0/2N$ and $t(p)$ and assume that $2N$ is large and that therefore $X_1/2N - p$ is small. Then using Taylor's theorem:

$$\begin{aligned} t(p) &= \mathbb{E}(t(p + \delta p)) + 1 \\ &\approx \mathbb{E}\left(t(p) + \delta p t'(p) + \frac{1}{2}\delta p^2 t''(p)\right) + 1 \\ &= t(p) + \mathbb{E}(\delta p)t'(p) + \frac{1}{2}\mathbb{E}(\delta p^2)t''(p) + 1 \end{aligned}$$

By Equation (3.2) we have $\mathbb{E}(\delta p) = 0$, and using Equation(3.3) we have

$$\mathbb{E}(\delta p^2) = \mathbb{E}((p_1 - p)^2) = \mathbb{V}\mathrm{ar}(p_1) = \frac{p(1-p)}{2N}$$

Thus, we find that

$$p(1-p)t''(p) \approx -4N$$

Solving this differential equation with boundary conditions $t(0) = 0$ and $t(1) = 0$ we find

$$\begin{aligned} t(p) &\approx -4N \iint \frac{1}{x(1-x)}\mathrm{dx} \\ &= -4N\left(p\log p + (1-p)\log(1-p)\right) \end{aligned} \tag{3.4}$$

as required.

In the case $i = 1$, so that $p = 1/2N$, which is the appropriate value if A was a unique allele due to a new mutation in an otherwise aa population, we find that

$$t(\frac{1}{2N}) \approx 2 + 2\log(2N) \text{ generations}$$

See question 4 on the examples sheet.

When $p = \frac{1}{2}$ we find

$$t(\frac{1}{2}) \approx 2.8N \text{ generations}$$

So for a human population of size $N = 10^6$ individuals, with average generation time of 20 years, we can expect to wait 56 million years for a gene that is initially present in the population to become fixed.

Put in this context we can see that genetic drift is only a very weak driver of genetic change for large populations. This also suggests why the Hardy-Weinberg law is useful, despite assuming infinite population size. We've shown that genetic drift has a timescale of $2N$ generations, whereas random mating has a time scale of only one or two generations. Because these two forces have such different time scales, they don't usually interact in an interesting way. In any particular generation, the population will appear to be in HWE. The deviation of the frequency of a genotype from the HW value will be no more than about $1/(2N)$, which isn't a measureable deviation.

See question 6 on the examples sheet.

3.2 Mutation

Given that genetic drift acts to reduce variation, why aren't all populations completely homogenous? The reason is mutation. In the process of replication and the production of gametes, errors can occur in the new DNA sequences leading to new genes. These errors are the source of all genetic variation.

Mutations can be beneficial, deleterious, or neutral (ie no selective effect). As we haven't yet encountered selection,

we focus on neutral mutations. The 'Neutral Theory' of genetics (proposed by Kimura) claims that most DNA differences between alleles within a population are due to neutral mutation.

3.2.1 Infinite allele model

Lets assume that all mutations result in a unique allele - ie one that is different by state from all other alleles that have ever existed. This is called the infinite allele model.

Exercise: How many different alleles are one mutational step away from an allele at a locus that is 3000 base pairs long? How many are two mutational steps away?

Let u be the mutation rate for a particular locus, ie, the probability that an allele in an offspring is different from the allele it was derived from in the parent. This means that mutations in the population of $2N$ haploid individuals accumulate at rate $2Nu$. This is counter-acted by drift which removes variation at rate $1/2N$. The first question to ask, is what is the expected number of alleles in the population at equilibrium?

The value of $\mathcal{G}$ after one round of random mating with drift and mutation will be

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

By assuming that u is small (typical values for u are in the range 10^{-5} to 10^{-10}), we can approximate $(1-u)^2$ by $1-2u$ to find that

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

If we now substitute $\mathcal{H} = 1 - \mathcal{G}$ and rearrange we find that

$$\mathcal{H}' \approx \left(1 - \frac{1}{2N}\right)\mathcal{H} + 2u(1 - \mathcal{H}) \tag{3.5}$$

which gives the change in $\mathcal{H}$ in a single generation to be

$$\Delta\mathcal{H} \approx -\frac{1}{2N}\mathcal{H} + 2u(1 - \mathcal{H}) \tag{3.6}$$

At equilibrium ($\Delta\mathcal{H} = 0$), we find that

$$\mathcal{H} = \frac{4Nu}{1 + 4Nu}$$

and

$$\mathcal{G} = \frac{1}{1 + 4Nu}.$$

We can break Equation (3.6) down further, by writing

$$\Delta\mathcal{H} = \Delta^{drift}\mathcal{H} + \Delta^{mut.}\mathcal{H}$$

where

$$\Delta^{drift}\mathcal{H} = -\frac{1}{2N}\mathcal{H}$$

is the change due to genetic drift, and

$$\Delta^{mut.}\mathcal{H} = 2u(1 - \mathcal{H})$$

is the change due to mutation.

Equilibrium occurs when the rate of change due to mutation equals rate of change due to drift, ie, when $\Delta^{mut.}\mathcal{H} = -\Delta^{drift}\mathcal{H}$. This equilibrium is only interesting when $4Nu$ is moderate in magnitude. If $4Nu$ is very small, then genetic drift dominates and all genetic variation is eliminated. Conversely, if $4Nu$ is large, then mutation dominates and all alleles in the population are different.

3.2.2 Wright-Fisher model of mutation

We can extend the Wright-Fisher model so that it also incorporates mutation, as well as genetic drift. Suppose that gene A mutates to a at rate u, but that there is no mutation from a to A. Then we can replace the model 3.1 by

$$P_{rx} = \mathbb{P}(X_{t+1} = x | X_t = r) = \binom{2N}{x} (\psi_r)^x (1 - \psi_r)^{2N-x} \tag{3.7}$$

where

$$\psi_r = \frac{r(1-u)}{2N}$$

Here, loss of A is certain.

If we suppose further that a mutates to A at rate v, then we can use

$$\psi_r = \frac{r(1-u) + (2N-r)v}{2N}$$

You will be asked to simulate from these two models on the examples sheet.

X_t is still a Markov chain, but now there are no absorbing states. We know that any ergodic irreducible Markov chain will converge to a stationary distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_{2N})$, where π_i is the probability there are i alleles of type A at equilibrium. Because π is the stationary distribution, we can write

$$\pi = \pi P$$

where P is the transition matrix given by Equation (3.7).

Lets consider the mean number of A alleles when the model is at equilibrium, namely

$$\mu = \sum_{i=0}^{2N} i\pi_i$$

We can write this as

$$\begin{aligned}
\mu &= \pi\xi \text{ where } \xi = (0, 1, \ldots, 2N)^T \\
&= \pi P\xi \\
&= \sum_i \pi_i \left(\sum_j P_{ij}\xi_j \right) \\
&= \sum_i \pi_i \left(\sum_j j \binom{2N}{j} (\psi_i)^j (1-\psi_i)^{2N-j} \right) \\
&= \sum_i \pi_i (2N\psi_i) \\
&= \sum_i (i(1-u) + (2N-i)v)\, \pi_i \\
&= \mu(1-u) + v(2N-\mu)
\end{aligned}$$

Hence, if we solve for μ, we find that

$$\mu = \frac{2Nv}{u+v}$$

Chapter 4

Selection

Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

Selection is the evolutionary force most responsible for adaptation to the environment. Selection acts when alleles have different fitnesses, and can occur in a variety of ways. For example:

- Unfair meiosis, for example due to sperm or pollen competition
- Fertility selection - the number of offspring produced may depend on maternal or paternal genotype.
- Viability selection - survival from zygote to adult may depend on genotype.
- Sexual selection - some individuals may be more successful at finding mates than others. Since females are typically the limiting sex (Bateman's principle) the differences typically arise either as a result of male-male competition or female choice.

We will focus solely on viability selection, partly because its the most important and illustrates all of the basic principles, and also because its easier to understand mathematically than the others.

4.1 Viability selection

Consider an autosomal locus in a hermaphroditic species. Viability selection acts between conception and sexual maturity. Hardy-Weinberg equilibrium holds only at the moment of conception, i.e., in zygote frequency:

Freq in parental gen.	$\rightarrow$	Freq in zygotes	$\rightarrow$	Freq in adults	$\rightarrow$	Freq in zygotes
p		p		p'		p'

Suppose that the probability that a zygote survives to adulthood is determined by its genotype. Let w_{11}, w_{12} and w_{22} denote the probabilities of survival (called the viabilities) of zygotes of genotype A_1A_1, A_1A_2 and A_2A_2 respectively.

Then we can see that

$$\text{freq. after selection} = \text{zygote freq.} \times \text{viability}$$

This gives the following frequency table

Genotype:	A_1A_1	A_1A_2	A_2A_2
Freq. in zygotes:	p^2	$2pq$	q^2
Viability:	w_{11}	w_{12}	w_{22}
Freq. after selection:	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$

where $\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$ is the constant of proportionality, and is called the *mean fitness of the population* by population geneticists.

We usually convert viabilities to relative fitnesses, as only the relative size matters. We do this by dividing the viabilities by w_{22}.

Genotype:	A_1A_1	A_1A_2	A_2A_2
Freq. in zygotes:	p^2	$2pq$	q^2
Relative fitness:	$1+s$	$1+hs$	1
Freq. after selection:	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$

We call s the *selection coefficient*. It is a measure of the fitness of A_1A_1 relative to A_2A_2. Its sign determines which homozygote is fitter.

We call h the *heterozygous effect*. It is a measure of the fitness of the heterozygote relative to the selective difference between the two homozygotes. We can think of h as being a measure of dominance.

$h = 0$	A_2 dominant, A_1 recessive
$h = 1$	A_1 dominant, A_2 recessive
$0 < h < 1$	incomplete dominance
$h < 0$	under-dominance
$h > 1$	over-dominance.

Complete dominance rarely occurs. Even 'recessive' lethals in human populations are now thought to be cases of incomplete dominance with s large and h small but greater than zero. Complete dominance does occur for morphological traits, but mainly in traits that don't affect fitness.

If we let p denote the frequency of allele A_1 in the first generation, and p' the frequency in the next generation, then we can show that

$$\Delta p = p' - p = \frac{sp(1-p)(p + h(1-2p))}{sp(p + 2h(1-p)) + 1}$$

which describes the deterministic behaviour of the population.

If we assume s and sh are small (typically less than 1%), then we can write

$$\Delta p \approx sp(1-p)(p+h(1-2p))$$

and if we measure time in units of one generation, we can approximate this equation by

$$\frac{\mathrm{dp}}{\mathrm{dt}} \approx sp(1-p)(p+h(1-2p)) \tag{4.1}$$

Let $t(p_1, p_2)$ be the time required for the frequency of A_1 to move from some value p_1 to some other value p_2, then we can see that

$$t(p_1,p_2) = \int_{p_1}^{p_2} \frac{1}{sp(1-p)(p+h(1-2p))}\mathrm{dp}$$

assuming that the frequency will eventually reach p_2.

On the example sheet, you will be asked to characterize the types of dominance behaviour that arise from Equation (4.1).

For example, suppose that $s > sh > 0$. Then it is clear from Equation (4.1) that the frequency of A_1 slowly approaches 1. However, there is a $(1-p)$ term in the denominator in the integrand, so as p approaches 1 the rate of increase slows and the time required even for small changes in p will be large.

The special case of $h = 1/2$ is of particular interest, as this represents additive fitness (the heterozygote is exactly intermediate between the two homozygotes) which is common for many alleles with very small values of s. Q3 on the example sheet asks you to find the number of generations spent in various frequency ranges for this case.

4.2 Selection and drift

We now consider the interation of selection and genetic drift. In an infinite population, an allele with a selective advantage will eventually become fixed in the population.

However, in finite populations there is a good chance that new selectively advantageous mutations will be lost, due to drift.

Natural selection is always a very weak force for rare alleles, the behaviour of which is determined by both drift and selection. For common alleles, the dynamics are mainly determined by natural selection (as long as $s >> 1/2N$).

Lets consider a haploid population with $2N$ individuals for which there is two alleles, A and a, with A enjoying a selective advantage s over wild type a. Then we aim to find the probability of fixation of allele A given that its current frequency is p, which we donote by $\pi(p)$.

We will show that

$$\pi(p) \approx \frac{1 - \mathrm{e}^{-4\mathrm{Nsp}}}{1 - \mathrm{e}^{-4\mathrm{Ns}}} \tag{4.2}$$

Proof: Consider the change of frequency going from p to $p + \delta p$ in a single generation. By the nature of Mendelian inheritance, the change in p is Markovian, and so we can write $f(p, p + \delta p)$ for the probability of a change from p to $p + \delta p$ in a single generation. We can write a backward equation for the probability of fixation:

$$\begin{aligned}
\pi(p) &= \int f(p, p + \delta p)\pi(p + \delta p)\mathrm{d}\delta\mathrm{p} \\
&\approx \int f(p, p + \delta p)\left(\pi(p) + \delta p\pi'(p) + \frac{1}{2}\delta p^2\pi''(p)\right)\mathrm{d}\delta\mathrm{p} \\
&= \pi(p) + m(p)\pi'(p) + \frac{1}{2}v(p)\pi''(p)
\end{aligned}$$

where $m(p) = \int \delta p f(p, p+\delta p)\mathrm{d}\delta\mathrm{p}$ is the mean change in p given initial frequency p, and $v(p) = \int \delta p^2 f(p, p + \delta p)\mathrm{d}\delta\mathrm{p}$ which approximately equals the variance of the change in p (if $m(p)$ is small).

Rearrange this last equation, and setting $f(p) = \pi'(p)$ gives the differential equation

$$f'(p) + \frac{2m(p)}{v(p)}f(p) = 0$$

We can solve this using the integrating factor

$$\exp(B(p)) \text{ where } B(p) = 2\int_0^p \frac{m(x)}{v(x)}\mathrm{dx} \tag{4.3}$$

giving

$$\frac{\mathrm{d}}{\mathrm{dp}}\left(f(p)e^{B(p)}\right) = 0$$

This implies

$$f(p) = c_1\mathrm{e}^{-\mathrm{B(p)}}$$
$$\pi(p) = c_1\int_0^p \mathrm{e}^{-\mathrm{B(y)}}\mathrm{dy} + \mathrm{c}_2$$

We have boundary conditions $\pi(0) = 0$ and $\pi(1) = 1$. The first condition implies $c_2 = 0$ and the second that

$$c_1 = \frac{1}{\int_0^1 e^{-B(y)}\mathrm{dy}}$$

giving

$$\pi(p) = \frac{\int_0^p e^{-B(y)}\mathrm{dy}}{\int_0^1 e^{-B(y)}\mathrm{dy}} \tag{4.4}$$

To find $m(p)$ and $v(p)$ consider the following table:

Genotype	A	a
Freq before selection	p	q
Fitness	1+s	1
Freq. after selection	$\frac{p(1+s)}{1+ps}$	$\frac{1-p}{1+ps}$

If X' denotes the number of A alleles after one round of mating and selection, then the Wright-Fisher model says that

$$X' \sim \text{Bin}\left(2N, \frac{p(1+s)}{1+ps}\right)$$

Then $m(p) = \mathbb{E}(\delta p) = \mathbb{E}(\frac{X'}{2N} - p)$ where X is the current number of A individuals. Thus

$$m(p) = \frac{p(1+s)}{1+ps} - p = \frac{spq}{1+ps} \approx spq$$

as long as s is small.

We can see that

$$\begin{aligned} v(p) &= \mathbb{V}\mathrm{ar}(\frac{X'}{2N} - p) \\ &= \frac{1}{4N^2}\mathbb{V}\mathrm{ar}(X') \\ &= \frac{1}{4N^2}2N \cdot \frac{p(1+s)}{1+ps} \cdot \frac{1-p}{1+ps} \\ &= \frac{1}{2N}pq\frac{1+s}{(1+ps)^2} \\ &\approx \frac{pq}{2N} \end{aligned}$$

Substituting $m(p)$ and $v(p)$ into Equation (4.3) gives

$$B(y) = 2\int_0^y \frac{pqs}{pq/2N}\mathrm{dx} = 4\mathrm{Nsy}$$

and substituting this into Equation (4.4) gives

$$\begin{aligned} \pi(p) &= \frac{\int_0^p e^{-4Nsy}\mathrm{dy}}{\int_0^1 e^{-4Nsy}\mathrm{dy}} \\ &= \frac{1-\mathrm{e}^{-4\mathrm{Nsp}}}{1-\mathrm{e}^{-4\mathrm{Ns}}} \end{aligned}$$

as required.

The most important application of Equation (4.2) is for finding the fixation probability of a new mutation, ie., when $p = 1/2N$, which is

$$\pi(\frac{1}{2N}) = \frac{1-e^{-2s}}{1-e^{-4Ns}}$$

Now if s is small, then $e^{-2s} \approx 1-2s$ and if $4Ns$ is large, then $e^{-4Ns} \approx 0$, giving

$$\pi(\frac{1}{2N}) \approx 2s$$

In other words, the probability of fixation for a new mutation is twice its selective advantage.

If we are considering a diploid population, where Aa heterozygotes have selective advantage $1+hs$ relative to wild type aa, then we simply replace s by sh, to find that the probability of fixation for a new mutation is $2sh$. So for

example, a new mutation with a 1% advantage when heterzygous, $hs = 0.01$, has only a 2% chance of fixation! A 1% advantage is rather strong selection. Once the allele is at all common, selection will overwhelm drift. For example, if $N = 10^6$ and the frequency of the allele is 0.01% ($p = 0.0001$), we get a probability of 0.86 for fixation. But there is still a 98% chance the mutation will be lost initially.

Chapter 5

Nonrandom Mating

Please email corrections and suggestions to
r.d.wilkinson@nottingham.ac.uk

In the previous two sections we have explored what happens to the Hardy-Weinberg equilibrium when we violate the following assumptions:

- infinite poplation size
- no mutation
- no selection

In this section we will violate the final major assumption, namely that individuals mate at random with respect to their genotype.

There are various ways in which individuals do not mate at random with respect to their genotype:

- Assortative mating - the tendency to mate with phenotypes similar to your own, e.g., IQ or attractiveness in humans
- Disassortative mating - the tendancy to mate with phenotypes that are different from your own - common in plants, also sex differences in animals.
- Inbreeding, including self-fertilization in plants, sib-mating, first-cousing mating, parent-offspring mating etc.

- Population subdivision - e.g., geographic separation.

We will only consider inbreeding and population subdivision, as assortative mating is similar.

The main outcome of non-random mating is increased levels of homozygosity. Homozygosity can lead to reduced viability and fitness, and so can reduce the health of a population.

5.1 Generalized Hardy-Weinberg

We introduce a new parameter, F, which we will give various interpretations throughout this section. However, we will see that for a range of definitions it leads to the same set of equations for gene frequencies.

To begin with, let

$F =$ probability of homozygosity due to special circumstances (PHSC)

These special circumstances might be inbreeding, subdivision, self-fertilisation etc. We can then derive the generalized HW equations.

First consider the probability of a individual having genotype AA. If we pick a random individual, and then pick one of its two alleles, the probability it is A is p. The probability that the other allele is also A is then

$$F + p(1 - F)$$

This breaks down into probability F that homozygosity is due to special circumstances, or special circumstances don't apply (with probability $1-F$) so that random-mating occurs, in which case the probability of an A is p. So the probability of a randomly choosen individual being AA is $pF + p^2(1 - F)$.

Similarly, the probability of genotype aa can be shown to be $q^2(1 - F) + qF$ where $q = 1 - p$. The probability of heterzogosity is found by subtracting the other probabilities from 1. Thus, the generalised HW frequencies are

Genotype	AA	Aa	aa
Frequency	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

Note that if $F = 0$ we recover the usual HW frequencies. We can see that because $0 \leq F \leq 1$, there will be an excess of homozygotes when compared to the usual HW frequencies. If $F \leq 0$, then there will be an excess of heterozygotes. Although we don't allow $F < 0$ when F=PHSC, under different interpretations of F, we might allow a negative value.

5.2 Inbreeding

5.2.1 Identity by descent

In order to quantify inbreeding, we need a quantitative measure of the relationship between relatives. Two alleles at a single locus are **identical by descent** (ibd) if they are identical copies of the same allele in some earlier generation, i.e., both are copies that arose by DNA replication from the same ancestral sequence without any intervening mutation. Note that identical by type is not the same as being identical by descent.

5.2.2 Coefficient of kinship

One useful measure of relationship, is the **coefficient of kinship**. Suppose there are two individuals, U and V, then the coefficient of kinship f_{uv}, is the probability that two alleles, one from U and one from V, are identical by descent.

Parent-offspring

What is the coefficient of kinship between a parent and their offspring? We begin by choosing an allele at random from the parent. The probability this is passed on to the offspring is $1/2$. We then pick a random allele from the

offspring. The probability that this came from the parent is also $1/2$. Hence

$$\begin{aligned} f_{PO} &= \mathbb{P}(ibd) \\ &= \mathbb{P}(\text{allele chosen in parent is passed to offspring}) \\ &\quad \times \mathbb{P}(\text{allele choosen in offspring comes from this parent}) \\ &= \frac{1}{4} \end{aligned}$$

Pictures can help.

Full sibs

First draw a picture!

Let U denote one sibling and V the other. Then

$$\begin{aligned} f_{FS} &= \mathbb{P}(ibd) \\ &= \mathbb{P}(\text{U \& V inherit same maternal allele}) \\ &\quad \times \mathbb{P}(\text{maternally inherited allele choosen in U \& V}|\text{same maternal allele inherited}) \\ &+ \mathbb{P}(\text{U \& V inherit same paternal allele}) \\ &\quad \times \mathbb{P}(\text{paternally inherited allele choosen in U \& V}|\text{same paternal allele inherited}) \\ &= \frac{1}{2}\frac{1}{4} + \frac{1}{2}\frac{1}{4} \\ &= \frac{1}{4} \end{aligned}$$

On the example sheet you will be asked to calculate the coefficient of kinship for half-sibs and first cousins.

5.2.3 More information

The fact that the coefficient of kinship is the same for parent-offspring and full-sibs points out the deficiency of using a single number to measure genetic relatedness. A parent and its offspring always have exactly one allele each that are ibd. Full-sibs, however, may have either zero, one or two pairs of alleles that are ibd (recall f is the probability of a random pair being ibd).

The most complete measure of relatedness is the set of probabilities of sharing 0, 1, or 2 pairs of ibd alleles. The following table gives some of these probabilities

Relationship	p_0	p_1	p_2
Parent-offspring	0	1	0
Full sibs	1/4	1/2	1/4
Half-sibs			
First cousins			

For full sibs, the probability they share two ibd alleles can be found as follows. Pick one of the alleles in individual U. The probability that an allele ibd with this allele is present in V is $1/2$. Now pick the other allele in U. As it necessarily came from the other parent, its chance of having an ibd allele in V is independent of the history of the first allele, and so this probability is also $1/2$. Thus, the probability of full sibs having two pairs of ibd alleles is $p_2 = 1/4$. We can reason similarly for p_0, and then obtain p_1 by subtraction.

Exercise: Complete the table for half-sibs and first cousins.

Given the values in the table above, the coefficient of kinship may be written in terms of the p_i as

$$f_{uv} = \frac{1}{4}p_1 + \frac{1}{2}p_2$$

The mean number of shared alleles is

$$\bar{p} = p_1 + 2p_2$$

and $p = \frac{1}{2}\bar{p}$ is called the coefficient of relatedness. Note that $f_{uv} = \frac{1}{2}p$.

5.2.4 Mating between relatives

Inbreeding occurs when an individual mates with a relative. The level of inbreeding is measured by the inbreeding coefficient F_I, which is the probability that two alleles in an individual are ibd. As one of the alleles comes from one parent and the other from the other parent, the inbreeding coefficient is just the coefficient of kinship of its parents $F_I = f_{xy}$.

If we calculate the genotype frequencies expected from mating between relatives, we find exactly the equations found for mating with special circumstances above. For

example, consider the frequency of AA genotypes. An individual could be AA for two reasons - either both alleles are identical by descent, or they are not but are both A due to random mating.

$$\begin{aligned}\mathbb{P}(AA) &= \mathbb{P}(AA|\text{ ibd })\mathbb{P}(\text{ ibd }) + \mathbb{P}(AA|\text{ not ibd })\mathbb{P}(\text{ not ibd }) \\ &= pF + p^2(1-F)\end{aligned}$$

which is exactly the expression found previously. The other two frequencies also turn out to be the same.

5.2.5 Inbreeding coefficient for the population

In general we will not be able to analytically calculate the inbreeding coefficient F for the population, as populations don't usually inbreed in a regular way. Instead, we are forced to estimate F for the entire population. One way to do this is to rearrange the equation for the heterozygosity $x_{12} = 2pq(1-F)$. This gives

$$F = 1 - \frac{x_{12}}{2pq}$$

If we assume that we observe the heterozygosity without error, and note that $2pq$ is the expected heterozygosity under Hardy-Weinberg, then we get another definition for F

$$F = 1 - \frac{\text{Observed heterozygosity}}{\text{Expected heterozygosity}}$$

If F is defined this way it is usually called the *population inbreeding coefficient*.

5.2.6 Self-fertilization

One particular form of inbreeding is self-fertilization. Many plants reproduce by a mixture of outcrossing and self-fertilization, in otherwords, they reproduce by a mixture of selfing and random mating.

Let σ be the fraction of progeny produced through self-fertilization. Let x_{AA} be the frequency of AA genotypes

in the current generation, and let x'_{AA} be the frequency in the next generation. Then it is easy to show that

$$\begin{aligned} x'_{AA} &= p^2(1-\sigma) + (x_{AA} + \frac{x_{Aa}}{4})\sigma \\ x'_{Aa} &= 2pq(1-\sigma) + \sigma\frac{x_{Aa}}{2} \\ x'_{aa} &= q^2(1-\sigma) + (x_{aa} + \frac{x_{Aa}}{4})\sigma \end{aligned}$$

Exercise: Verify that the allele frequencies don't change between parents and offspring.

Because homozygous parents can always have heterozygous offspring (when they outcross), heterozygotes are never completely eliminated. We can solve for the equilibrium frequency of heterzygotes:

$$\begin{aligned} \hat{x}_{Aa} &= 2pq(1-\sigma) + \sigma\frac{\hat{x}_{Aa}}{2} \\ &= \frac{2pq(1-\sigma)}{1-\sigma/2} \end{aligned}$$

and we can find similar expressions for $\hat{x}_{AA}$ and $\hat{x}_{aa}$. If we now let

$$F = \frac{\sigma}{2-\sigma}$$

(another definition for F), we then find

$$\hat{x}_{Aa} = 2pq(1-F)$$

as before, thus showing that this definition of F is equivalent to the others in this case! The same equations can also be derived for the frequency of $\hat{x_{aa}}$ and $\hat{x_{AA}}$.

Exercise: Suppose you observed without error, the following genotype frequencies in a plant species that engages in mixed selfing and outcrossing:

Genotype:	AA	Aa	aa
Frequency:	0.828	0.144	0.028

What is the frequency of selfing, σ, if the population is at equilibrium?

5.3 Estimating p and F

How do we estimate values of p and F? We could use the EM algorithm as before (indeed, the example sheet asks you to do just that), but instead this time we'll use a multivariate Newton-Raphson method, as seen previously.

Using second derivatives, we have a multivariate NR method. From point $\boldsymbol{\theta}^0 = (p_0, F_0)^T$ we move to point

$$\boldsymbol{\theta}' = \boldsymbol{\theta}^0 + \boldsymbol{I}(\boldsymbol{\theta}^0)^{-1} S(\boldsymbol{\theta}^0),$$

where $S(\theta) = \nabla l(\boldsymbol{\theta})$ is the gradient vector $= \left(\frac{\partial l}{\partial p}, \frac{\partial l}{\partial F}\right)^\top$ of the log-likelihood $l(\boldsymbol{\theta})$, usually called the score function in statistics.

$\boldsymbol{I}(\boldsymbol{\theta}^0)$ is the $p \times p$ negative Hessian matrix with (i, j) element $-\dfrac{\partial^2 l}{\partial \theta_i \partial \theta_j}$ usually called the observed information matrix.

This method is *very* fast if $\boldsymbol{\theta}^0$ starts sufficiently close to a solution. However it does require first and second derivatives and inversion of a matrix, so can be costly in terms of time spent deriving equations.

Exercise: Find p and F for a population of size 1000 for which the genotype frequences of AA, Aa and aa are 0.056, 0.288, and 0.656 respectively.

Lets first begin by writing down the likelihood. As before, assuming that the model is true, we find that the frequencies have a multinomial distribution

$$\mathbb{P}(n_{AA}, n_{Aa}, n_{aa}|p, F) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \left(pF + p^2(1-F)\right)^{n_{AA}} \times (2p(1-p)(1-F))^{n_{Aa}} \left((1-p)F + (1-p)^2(1-F)\right)^{n_{aa}}$$

If we apply NR starting from $p = 0.8$ and $F = 0.8$, we go through the following iterations

Iteration	p	F
0	0.8	0.8
1	0.6121254	0.6760076
2	0.2951159	0.4644971
3	0.1694039	0.1682794
4	0.19645003	0.09939278
5	0.1999503	0.0999270
7	0.19999999	0.09999997
8	0.2	0.1
9	0.2	0.1

quickly converging to the maximum likelihood estimate $\hat{p} = 0.2, \hat{F} = 0.1$. The Newton-Raphson method automatically provides an estimate of the covariance matrix of the MLE via the inverse of the observed or expected information matrix, evaluated at the MLE. Therefore it is possible to quantify the precision of the maximum likelihood estimate. In this case, we find

$$I_0^{-1}(\hat{\theta}) = \begin{pmatrix} 8.8e-05 & 0.000027000 \\ 2.7e-05 & 0.001182375 \end{pmatrix}$$

This gives $se(\hat{p}) = 0.0094$ and $se(\hat{F}) = 0.034$.

R code for this implementation (including algebraic calculation of derivatives) is given on the course website.

See question 4 on the examples sheet.