Population Genetics

Tutorial

Peter Pfaffelhuber, Pleuni Pennings,
and Joachim Hermisson

February 2, 2009

University of Vienna
Mathematics Department
Nordbergsrtaße 15
1090 Vienna, Austria

Preface

This tutorial was written for the course *Population Genetics Computer Lab* given at the Veterinary Medical University of Vienna in February 2008 and 2009. It consists of nine sections with lectures and computerlabs.

The course was taught as part of an intensive training course for incoming students of the PhD program in Population Genetics. It is designed for graduate students with diverse backgrounds, including biologists, bio-informaticians, and mathematicians and equally diverse plans for their PhD thesis. In particular, the cause addresses theoreticians and empiricists. Although a basic understanding of genetics, statistics and modeling is definitely useful, it is not a strict requirement. Short introductions to each of these subjects is provided in the course.

The aim is to introduce population genetic methods in a combined approach, from the data side as well as from a modelling point of view. On the one hand, we explain the mathematical concepts that are needed to understand basic population genetic models. On the other hand, it is shown how these models can be used when they are applied to data. After following the course, students should have a basic understanding of the most prominent methods in molecular population genetics that are used to analyze data. Additional material and methods that reach far beyond the scope of this tutorial can be found in several textbooks, such as Durrett (2002), Ewens (2004), Gillespie (2004), Halliburton (2004), Hartl and Clark (2007), Hedrick (2005) or Nei (1987).

We are grateful to Tina Hambuch, Anna Thanukos and Montgomery Slatkin, who developed former versions of this course from which we profited a lot. Agnes Rettelbach helped us to fit the exercises to the needs of the students and Ulrike Feldmann was a great help with the `R`-package `labpopgen` which comes with this course. Toby Johnson kindly provided material that originally appeared in (Johnson, 2005) which can now be found in Sections 1 and 9.

Peter Pfaffelhuber, Pleuni Pennings, Joachim Hermisson

Contents

1 Polymorphism in DNA

DNA is now understood to be the material on which inheritance acts. Since the 1980s it is possible to obtain DNA sequences in an automated way. Already one round of classical sequencing - if properly executed and if the sequencer works well - gives up to 1000 bases of a DNA stretch. Automated sequencers can read from 48 to 96 such fragments in one run. The whole procedure takes about one to three hours. Most recently, a new generation of high-throughput sequencers has entered the stage. These sequencers produce usually (much) shorter reads, but can easily generate data from several 100 million nucleotides per day. As can be guessed from these numbers there is a flood of data generated by many labs all over the world. The main aim of this course is to give some hints how it is possible to make some sense out of these data, especially, if the sequences come from individuals of the same species.

1.1 The life cycle of DNA

The processing of DNA in a cell of an individual runs through different stages. Since a double strand of DNA only contains the instructions how it can be processed, it must be read and then the instructions have to be carried out. The first step is called transcription and results in one strand of RNA per gene. The begin and end of the transcription tract determine a gene[1]. DNA regions that are not transcribed are called intergenic. The initial transcript, or pre-mRNA, is further processed to excise *introns* and splice the *exons* together. DNA in exons is called *coding*, all other DNA (i.e. intergenic regions and introns) are *non-coding*. The resulting messenger or mRNA transcript is the template for the second main step of information processing, called translation. Translation results in proteins or polypeptides which the cell can use and process. During transcription exactly one DNA base is transcribed into one base of RNA, but in translation three bases of RNA encode an *amino acid*. The combinations of the three base pairs are called *codons* and the translation table of which codon gives which amino acid is the *genetic code*. There is a certain redundancy in this mechanism because there are $4^3 = 64$ different codons, but only 20 different amino acids. Certain amino acids are thus represented by more than one set of three RNA bases. In particular, the third basepair of the codon is often redundant (or *silent*), which means that the amino acid is already determined by the first two RNA bases.

As DNA is the material of genetic inheritance it must somehow be transferred from ancestor to descendant. Roughly speaking we can distinguish two reproduction mechanisms: sexual and asexual reproduction. Asexually reproducing individuals only have one parent. This means that the parent passes on its whole genetic material to the offspring. The main example for asexual reproduction is *binary fission* which occurs often in prokaryotes, such as bacteria. Another instance of asexual reproduction is *budding*, which is the process by which offspring is grown directly from the parent, or the use of somatic cell nuclear transfer

[1]The name *gene* is commonly used with several different meanings.

for reproductive cloning.

Different from asexual is sexual reproduction. Here all individuals have two parents. Each progeny receives a full genomic copy from each parent. That means that every individual has (usually) two copies of each chromosome, which both carry the instructions the individual would need to build its proteins. So there is an excess of information stored in the individual. When an individual passes on its genetic material to its offspring it only gives one set of chromosomes. Due to *recombination* during cell division, the genetic material that is transfered to a child is a mixture of the material coming from the parent's own mother and father. Since the child receives a set of chromosomes from both parents, it has two sets of chromosomes again. The reduction from a diploid set of chromosomes to a haploid one occurs during *meiosis*, the process when gametes are produced which have half of the number of chromosomes found in the organism's diploid cells. When the DNA from two different gametes is combined, a diploid zygote is produced that can develop into a new individual.

Both with asexual and sexual reproduction, *mutations* can accumulate in a genome. These are 'typos' made when the DNA is copied during cell division. We distinguish between point mutations and indels. A *point mutation* is a site in the genome where the exact base `A`, `C`, `G`, or `T`, is exchanged by another one. *Indels* (which stands as an abbreviation for *in*sertions and *del*etions) are mutations where some DNA bases are erroneously deleted from the genome or some others are inserted. Often we don't know whether a difference between two sequences is caused by an insertion or a deletion. The word *indel* is an easy way to say that one of the two must have happened.

As we want to analyze DNA data it is important to think about what data will look like. The dataset that we will look at in this section will be taken from two lines of the model organism *Drosophila melanogaster*. We will use `DNASP` to analyze the data. A *Drosophila* line is in fact a population of genetically almost identical individuals. By inbreeding a population will become more and more the same. This is very practical for sequencing experiments, because it means that you can keep the DNA you want to study, although you kill individuals that carry this DNA.

Exercise 1.1. Open `DNASP` and the file `055twolines.nex`. You will find two sequences. When you open the file a summary of your data is displayed. Do you see where in the genome your data comes from? You can get an easy overview of your data in clicking `Overview->Intraspecific Data`. How many differences due to point-mutations (usually called single nucleotide polymorphisms or SNPs) are there in the two sequences? Are there also indel polymorphisms? □

Alignments

The data you looked at were already nicely prepared to be loaded into `DNASP`. Usually they first must be *aligned*. This task is not trivial. Consider two homologous sequences

```
A T G C A T G C A T G C
A T C C G C T T G C
```

They are not identical so we need to think which mutational mechanisms can account for their differences. Since the second sequence is shorter than the first one, we already know that at least one indel must have taken place. An alignment of the two sequences is an arrangement of the two sequences such that homologous bases are in the same column. Since we only have our data from extant individuals we can never be sure about which bases are homologous. Therefore there exist many possible alignments. We could for example try to align the sequences by introducing many insertions and deletions and no point mutations, such as

```
A T - G C A T G C A - T G C
A T C - C - - G C - T T G C
```

This alignment contains six indels but no point mutations. Another possible alignment would be

```
A T G C A T G C A T G C
A T C C - - G C T T G C
```

where we have used two point mutations and one indel of length two. Which alignment you prefer depends on how likely you think point mutations are relative to indels. Usually, the way to decide what is the best alignment is by first deciding upon a scoring system for indels and point mutations. The scoring may be different for different lengths of indels and different point mutations. Events that happen often have a low score, events that are rare have a high score. Now it is possible to calculate a score for every possible alignment, because every alignment is a statement about the events that happened in the history of the sample. The alignment with the lowest score is considered the best.

Exercise 1.2. Align the two sequences using only indels. Repeat using only point mutations. Now find the alignment with the least number of mutations, given that point mutations and indels each equally likely.

```
A A T A G C A T A G C A C A C A
T A A A C A T A A C A C A C T A
```

□

1.2 Various kinds of data

Patterns of diversity can be studied for various kinds of data. You may compare DNA sequences of several species or you may study the diversity within a single species. Analyses concerned with reconstruction of phylogenetic trees fall into the former category. *Population genetics* deals mostly with variation within a species. Both fields overlap, however, and we will see (already in this section) that population genetics sometimes also uses comparisons among species.

So the most elementary thing to know about a data set is whether it comes from one or more than one species. But there are many more questions you can (should) ask when looking at any dataset containing DNA sequences. For example:

1. Are the sequences already aligned?
2. Are the data from one population or more than one?
3. Are the data from a sexually or asexually reproducing organism?
4. Are the sequences from coding or non-coding DNA?
5. Are the data from one or more loci?
6. Do we see all sites or only the variable ones (SNPs, indels, or both)?
7. Do we see all sequences or only the different ones?
8. Is the data from microsatellites?

A microsatellite is a short stretch of DNA that is repeated several times. It could for example be `ATATATATATAT`. A common mutation in a microsatellite is a change in the number of repeats of the short DNA stretch which is a special form of an indel. That is the reason why they are also called VNTRs ('variable number of tandem repeats'). They are usually found in non-coding DNA. The advantage of using them is that you do no need to sequence a whole stretch of DNA but only use electrophoresis to infer the number of repeats. The disadvantage is that they do not contain as much information as SNP data.

Exercise 1.3. Can you answer the above questions for the dataset you looked at in Exercise 1.1? □

The most important mechanisms that shape DNA sequence variation are: mutation, selection, recombination and genetic drift. We will start with mutation, as it is maybe the simplest mechanism and the one that is most obvious when one starts to look at data. The other mechanisms will be explained later. Their effects in isolation and combination will be made clear during the course.

1.3 Divergence and estimating mutation rate

Our little dataset `055twolines.nex` that we consider next consists of two sequences. The two sequences come from two populations of *Drosophila melanogaster*, one from Europe and the other one from Africa. They must have a most recent common ancestor (MRCA) sometime in the past. Looking at the data one sees that the sequences are not identical. As the common ancestor only had one DNA sequence this sequence must have changed somehow in the history between the MRCA and the individuals today.

An important idea for the interpretation of mutations is the idea of the *molecular clock*. It says that when a piece of genetic material passes from one generation to the next there is a constant probability - which we denote by μ - that a mutation occurs. The important word here is *constant*. Non-constant mutation rates would mean that there are times when

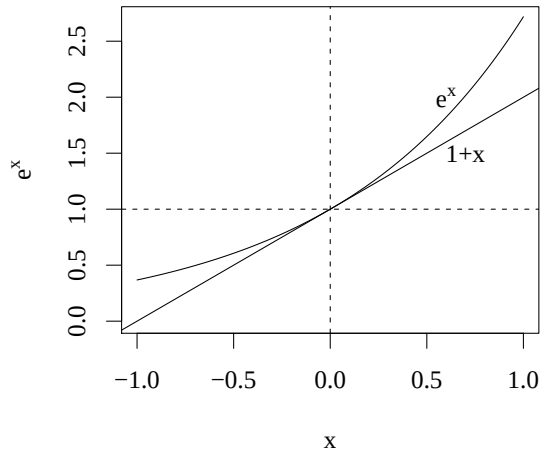


Figure 1.1: Curves of $1 + x$ and e^x are close together near $x = 0$.

more mutations will accumulate and times with fewer ones. Over larger evolutionary times, we know that mutation rates are not constant, but for now we will assume they are.

We have to be specific when we speak about the probability of mutation μ. It can either be the probability that a mutation occurs at a certain site (which would be the per site mutation rate) or on the scale of an entire locus (which would then be the locus wide mutation probability), and we can also consider the genome wide mutation rate. In the following it doesn't matter, which unit of the genome we consider.

If μ is the mutation probability per generation, $(1 - \mu)$ is the probability that no mutation occurs. Consequently,

$$\mathbf{P}[\text{no mutation for } t \text{ generations}] = (1 - \mu)^t.$$

is the probability that no mutation has occured in the past t generations in a line of descent. There is an approximation as long as μ is small compared to t.

Maths 1.1. *As long as x is small,*

$$1 + x \approx e^x, \qquad 1 - x \approx e^{-x}.$$

This can be seen by looking at the graph of the function $e^{\cdot}$ shown in Figure 1.1

By this approximation the probability that there is no mutation for t generations is

$$\mathbf{P}[\text{no mutation for } t \text{ generations}] \approx e^{-\mu t}.$$

The approximation can be used as long as μ is small, which is typically the case as long as we consider only a site or a small stretch of DNA.

We can also describe a *probability distribution* for the time until the next mutation on that specific ancestral lineage. Since probabilities and *random variables* will be frequently used in the course, we will first give a basic introduction into these concepts.

Maths 1.2. *A* random variable *(usually denoted by a capital letter) is an object which can take certain values with certain probabilities. The values can be either discrete or continuous. The probabilities are determined by its* distribution. *The probability that a random variable X takes a value x is denoted*

$$\mathbf{P}[X = x], \qquad \mathbf{P}[X \in dx]$$

for discrete and continuous random variables respectively. The (cumulative) distribution function of X is given by

$$F_X(x) := \mathbf{P}[X \leq x]$$

for each of the two cases. This function is increasing and eventually reaches 1. It uniquely determines the distribution of X.

In most cases a random variable has an expectation and a variance. They are given by

$$\begin{aligned}\mathbf{E}[X] &= \sum_x x\mathbf{P}[X = x],\\ \mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \sum_x (x - \mathbf{E}[X])^2\mathbf{P}[X = x]\end{aligned}$$

for discrete random variables and

$$\begin{aligned}\mathbf{E}[X] &= \int x\mathbf{P}[X \in dx],\\ \mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \int (x - \mathbf{E}[X])^2\mathbf{P}[X \in dx]\end{aligned}$$

for continuous ones.

In the above case we are dealing with the random variable

$$T := \text{ time to the next mutation.}$$

For its distribution we calculated already

$$\mathbf{P}[T \geq t] \approx e^{-\mu t}.$$

These probabilities belong to the *exponential distribution.*

Maths 1.3. *Let X be exponentially distributed with parameter λ. This means that*

$$\mathbf{P}[X \in dx] = \lambda e^{-\lambda x} dx, \qquad \mathbf{P}[X \geq x] = e^{-\lambda x}$$

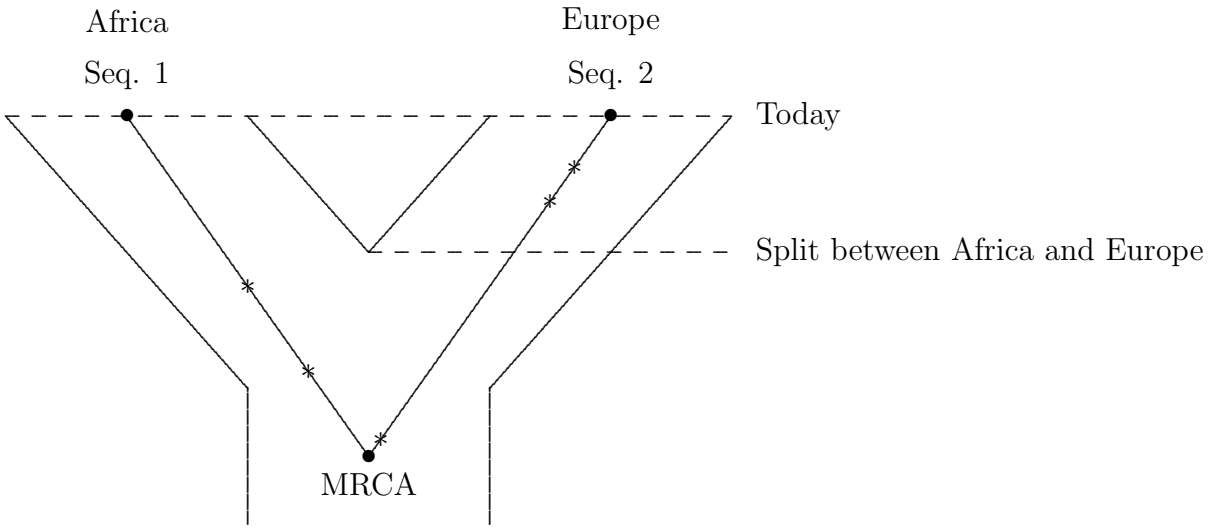


Figure 1.2: Mutations on ancestral lines in a sample of size 2, one from Europe and one from Africa

and

$$\mathbf{E}[X] = \lambda \int_0^\infty x e^{-\lambda x} dx = -\lambda \frac{d}{d\lambda} \int_0^\infty e^{-\lambda x} dx = \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda}$$
$$\mathbf{E}[X^2] = \lambda \int_0^\infty x^2 e^{-\lambda x} dx = -\lambda \frac{d^2}{d\lambda^2} \int_0^\infty e^{-\lambda x} dx = \lambda \frac{2}{\lambda^3} = \frac{2}{\lambda^2}$$
$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{1}{\lambda^2}.$$

Usually, the parameter λ is referred to as the rate *of the exponenital distribution.*

So T is approximately *exponentially distributed* with parameter μ. The expectation and variance of this waiting time are thus

$$\mathbf{E}[T] = \frac{1}{\mu}, \qquad \mathbf{Var}[T] = \frac{1}{\mu^2}.$$

Cosnider the setting in Figure 1.2. Assume we know the time since the two populations split was t generations ago. Since the common ancestor of the two lines can only be found before that time, we know that the two individuals must be separated by at least $2t$ generations. Now, lets assume they are seperated by exactly $2t$ generations. The time since the split between African and European *Drosophila* populations is not so long, approximately 10 KY, which we assume to be 100.000 generations. From this time we can estimate the mutation rate μ assuming a molecular clock. (Just to compare: the time since the split of humans and chimpanzees is about 5 MY, or 250.000 generations.)

Parameter estimation is a general mathematical concept and works in any quantitative setting.

Maths 1.4. *Given any model with a model parameter* $\bullet$ *a random variable* $\widehat{\bullet}$ *is called an estimator of* $\bullet$*. It is called an unbiased estimator if*

$$\mathbf{E}[\widehat{\bullet}] = \bullet$$

where $\mathbf{E}[.]$ *is the expectation with respect to the given model.*

Here $\bullet = \mu$, so we want to obtain an estimator for μ. Obviously we must base this estimator on D the number of polymorphic sites ot the *divergence* between the two populations. Let us first think the other way round. Tracing back our two lines for time t (and assume we know μ), K mutations have occured along the branches of the two descendant populations which has length $2t$. It is possible, however that two mutations hit the same site in the chromosome. If this is the case today we can only see the last mutant. Assuming that divergence is small enough such that to a good approximation each site is hit at most once by a mutation, we set $K = D$. As mutations occur at constant rate μ along the branches

$$\mathbf{E}[D] = \mathbf{E}[K] = 2\mu t. \tag{1.1}$$

This already gives a first unbiased estimator for μ, which is

$$\widehat{\mu} = \frac{D_t}{2t}.$$

However, if mutations hit sites that were already hit by a mutation, this messes these thoughts up. The longer the time since divergence or the higher the mutation rate the higher the chance of these double hits. So, if divergence is too big, the assumption of no double hits might be misleading. The so-called Jukes-Cantor corrections account for this effect.

Exercise 1.4. Look at your data. Assume the European and African lines of *Drosophila melanogaster* have separated 10,000 years (10KY) ago. What is your estimate for the mutation rate μ? Given that for the two populations the time of their split is not exactly the time the two individuals have a common ancestor, is your estimate for μ an over- or and underestimate (upper or lower bound)? □

Divergence between species

All of the above analysis also works for divergence between species. Divergence data are often summarized by the number of substitutions between each pair of species, such as those shown in Figure 1.3. The species tree for the three species is given in Figure 1.4.

A *relative rates test* is used to test whether there is a constant mutation rate in the different lineages in the tree. It is a test of the molecular clock. The lineages studied would be those leading to sequences A and B, and we use a third sequence C as an outgroup to A and B (Figure 1.4). If the molecular clock operates there must be as much divergence between A and C as between B and C.

	OW Monkey (B)	NW Monkey (C)
Human (A)	485 (0.072)	1201 (0.179)
OW Monkey(B)		1288 (0.192)

Figure 1.3: Number of pairwise differences (and fraction) for $l = 6724$bp of aligned η-globin pseudogene sequence. The OW (Old World) monkey is the rhesus macaque and the NW (New World) monkey is the white fronted capuchin.

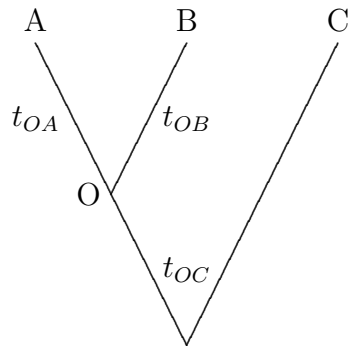


Figure 1.4: Tree topology for relative rates test

If the divergence is small then multiple hits can be ignored, and ancestral sequences can be reconstructed using parsimony,[2]. i.e. by minimizing the number of mutations on the lines to the MRCA. If the observed number of differences between three sequences A, B and C are k_{AB}, k_{AC} and k_{BC}, then the reconstructed number of substitutions since O, the common ancestor of A and B, can be computed as

$$k_{OA} = \frac{k_{AB} + k_{AC} - k_{BC}}{2}$$
$$k_{OB} = \frac{k_{AB} + k_{BC} - k_{AC}}{2}$$
$$k_{OC} = \frac{k_{AC} + k_{BC} - k_{AB}}{2}.$$

The reconstructed numbers of substitutions k_{OA} and k_{OB} can now be analyzed. Assume the rates on the branches OA and OB occur at the same rate. Then, given k_{AB} every mutation occurs on the branch OA with probability $\frac{1}{2}$. This leads us to binomial distributions.

[2]The principle of parsimony (also called Ockham's Razor) states that one should prefer the least complex explanation for an observation. In systematics, maximum parsimony is a cladistic "optimality criterion" based on the principle of parsimony. Under maximum parsimony, the preferred phylogenetic tree (or alignment) is the tree (or alignment) that requires the least number of evolutionary changes.

Maths 1.5. *If a random variable X is binomially distributed with parameters p and n this means that*

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

where k is between 0 and n. This is the probability when you do a random experiment which has two possible results n times you get one result (which has for one instance of the experiment a probability of p) exactly k times.

Note that because there must be some outcome of the experiment,

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Our success probability is $p = \frac{1}{2}$ and the number of experiments we do is k_{AB} because we place all mutations randomly on OA and OB. In our example

$$n = k_{AB} = 485, \quad k = k_{OA} = \frac{485 + 1201 - 1288}{2} = 199,$$
$$n - k = k_{OB} = \frac{485 + 1288 - 1201}{2} = 286.$$

We assume a constant rate of mutation and then test if the observed data is consistent with this model. The relative rates test asks for the probability that under the assumption of a constant rate observed data can be as or more different than the data we observed, i.e.

$$\mathbf{P}[K_{OA} \geq 286] + \mathbf{P}[K_{OA} \leq 199] = 2\mathbf{P}[K_{OA} \geq 286] = 9.1 \cdot 10^{-5},$$

so this probability is very small. This value is called the p-value in statistics. As it is very low we must reject the hypothesis that there was a molecular clock with a constant rate in both branches.

Exercise 1.5.

1. On your computer you find the program `R` which we will use frequently during the course. [3] `R` knows about the binomial distribution. Type `?dbinom` to find out about it. Can you repeat the calculation that led to the p-value of $9.1 \cdot 10^{-5}$ using `R` (or any other program if you prefer)?

2. Can you think of explanations why there was no clock with a constant rate in the above example? If you want you can use the internet to find explanations.

□

Exercise 1.6. Assume the homologous sequences of three species are

```
species 1:  ATG CGT ATA GCA TCG ATG CTT ATG GC
species 2:  ACG CCA CTG GCA ACC ATG CTA AAG GG
species 3:  ATG CGA CTA GCG TCC ATG CTA ATG GC
```

[3] A (very) short introduction to `R` and all procedures you need during the course can be found in Appendix A.

Which species do you assume to be most closely related? Count the number of differences in the sequences. Perform a relative rates test to see whether the assumption of rate constancy is justified. □

2 The Wright-Fisher model and the neutral theory

Although the main interest of population genetics is conceivably in natural selection, we will first assume that it is absent. Motoo Kimura developed the *neutral theory* in the 50s and 60s (see e.g. KIMURA, 1983). He famously pointed out that models without selection already explain much of the observed patterns of polymorphism within species and divergence between species. Today, the neutral theory is the standard null-model of population genetics. This means, if we want to make the case for selection, we usually do so by rejecting the neutral hypothesis. This makes understanding of neutral evolution key to all of population genetics.

Motoo Kimura, 1924–1994, published several important, highly mathematical papers on random genetic drift that impressed the few population geneticists who were able to understand them (most notably, Wright). In one paper, he extended Fisher's theory of natural selection to take into account factors such as dominance, epistasis and fluctuations in the natural environment. He set out to develop ways to use the new data pouring in from molecular biology to solve problems of population genetics. Using data on the variation among hemoglobins and cytochromes-c in a wide range of species, he calculated the evolutionary rates of these proteins. Extrapolating these rates to the entire genome, he concluded that there could not be strong enough selection pressures to drive such rapid evolution. He therefore decided that most evolution at the molecular level was the result of neutral processes like mutation and drift. Kimura spent the rest of his life advancing this idea, which came to be known as the "neutral theory of molecular evolution" (adapted from `http://hrst.mit.edu/groups/evolution`.)

2.1 The Wright-Fisher model

The Wright-Fisher model (named after Sewall Wright and Ronald A. Fisher) is the simplest population genetic model that we have. In this section you learn how this model is usually constructed and what its basic assumptions and characteristics are. We will introduce the model in its simplest shape, for a single locus in a haploid population of constant size. Under the assumption of *random mating* (or *panmixia*), a diploid population of size N can be described by the haploid model with size $2N$, if we just follow the lines of descent of all gene copies separately. (Technically, we need to allow for selfing with probability 1/N.)

Sewall Wright, 1889–1988; Wright's earliest studies included investigation of the effects of inbreeding and crossbreeding among guinea pigs, animals that he later used in studying the effects of gene action on coat and eye color, among other inherited characters. Along with the British scientists J.B.S. Haldane and R.A. Fisher, Wright was one of the scientists who developed a mathematical basis for evolutionary theory, using statistical techniques toward this end. He also originated a theory that could guide the use of inbreeding and crossbreeding in the improvement of livestock. Wright is perhaps best known for his concept of genetic drift (from Encyclopedia Britannica 2004).

Sir Ronald A. Fisher, 1890–1962, Fisher is well-known for both his work in statistics and genetics. His breeding experiments led to theories about gene dominance and fitness, published in The Genetical Theory of Natural Selection (1930). In 1933 Fisher became Galton Professor of Eugenics at University College, London. From 1943 to 1957 he was Balfour Professor of Genetics at Cambridge. He investigated the linkage of genes for different traits and developed methods of multivariate analysis to deal with such questions.
An even more important achievement was Fisher's invention of the analysis of variance, or ANOVA. This statistical procedure enabled experimentalists to answer several questions at once. Fisher's principal idea was to arrange an experiment as a set of partitioned subexperiments that differ from each other in one or more of the factors or treatments applied in them. By permitting differences in their outcome to be attributed to the different factors or combinations of factors by means of statistical analysis, these subexperiments constituted a notable advance over the prevailing procedure of varying only one factor at a time in an experiment. It was later found that the problems of multivariate analysis that Fisher had solved in his plant-breeding research are encountered in other scientific fields as well.
Fisher summed up his statistical work in his book Statistical Methods and Scientific Inference (1956). He was knighted in 1952 and spent the last years of his life conducting research in Australia (from Encyclopedia Britannica 2004).

As an example, imagine a small population of 5 diploid or 10 haploid individuals. Each of the haploids is represented by a circle. Ten circles represent the first generation (see Figure 2.1). In the neutral Wright-Fisher model, you obtain an offspring generation from a given parent generation by the following set of simple rules:

1. Since we assume a constant population, there will be 10 individuals in the offspring generation again.

2. Each individual from the offspring generation now picks a parent at random from the previous generation, and parent and child are linked by a line.

3. Each offspring inherites the genetic information of the parent.

The result for one generation is shown in Figure 2.2. After a couple of generations it will look like Figure 2.3(A). In (B) you see the *untangled version*. This picture shows the same process, except that the individuals have been shuffled a bit to avoid the mess of many lines crossing. The genealogical relationships are still the same, only the children of one parent are now put next to each other and close to the parent.

Almost all models in this course are versions of the Wright-Fisher model. We will describe later in this section how mutation can be built in, in Section 4 we will be concerned with inbreeding and substructured populations, in 5 we will allow for non-constant population size, in Section 6 we will extend the model to include recombination and finally in Section 7 we will deal with the necessary extensions for selection.

Figure 2.1: The 0th generation in a Wright-Fisher Model.

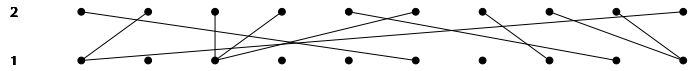


Figure 2.2: The first generation in a Wright-Fisher Model.

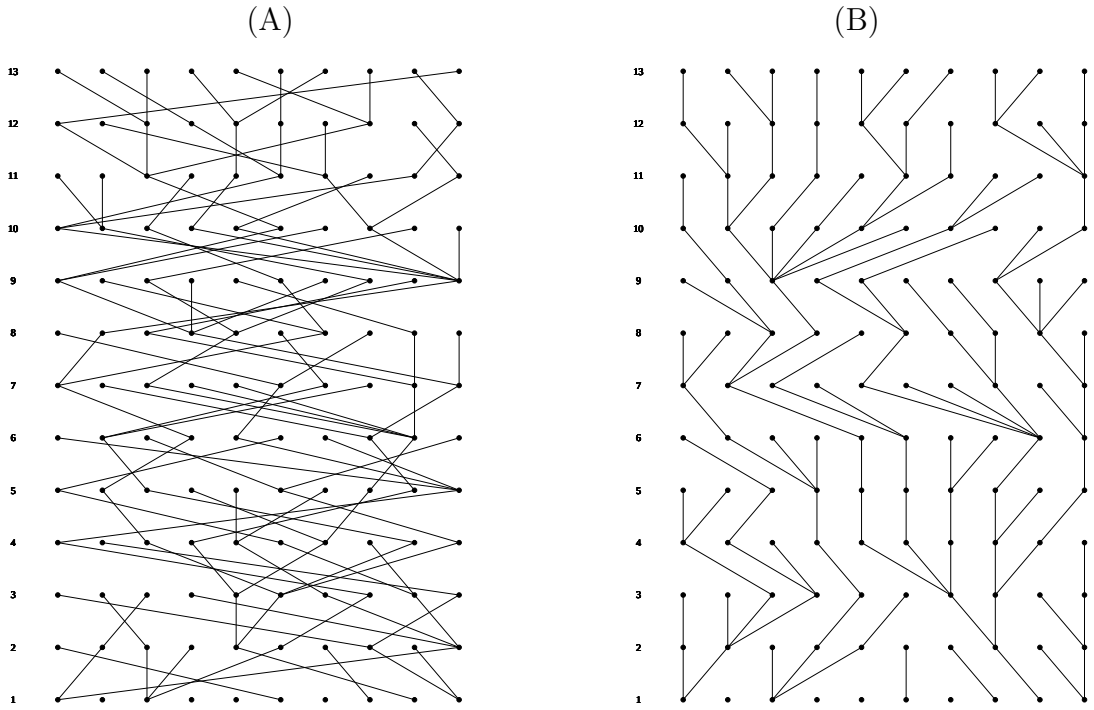


Figure 2.3: The tangled and untangled version of the Wright-Fisher Model after some generations.

Neutral evolution means that all individuals have the same *fitness*. Fitness, in population genetics, is a measure for the expected number of offspring. In the neutral Wright-Fisher model, equal fitness is implemented by equal probabilities for all individuals to be picked as a parent.

Each individual will therefore have $2N$ chances to become ancestor of the next generations and in each of these "trials" the chance that it is picked is $\frac{1}{2N}$. That means that the number of offspring of each individual is binomially distributed with parameters $p = \frac{1}{2N}$ and $n = 2N$ (see Maths 1.5). For a large population, n is large and p is small. In this limit, the binomial distribution can be approximated by the Poisson distribution.

Maths 2.1. *If a random variable X is binomially distributed with parameters n and p such that n is big and p is small, but $np = \lambda$ has a reasonable size, then*

$$\begin{aligned}\mathbf{P}[X = k] &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n \cdots (n-k+1)}{k!} p^k \big(1 - \tfrac{\lambda}{n}\big)^n / (1-p)^k \\ &\approx \frac{n^k}{k!} p^k e^{-\lambda} = e^{-\lambda} \frac{\lambda^k}{k!}.\end{aligned}$$

These are the weights of a Poisson distribution and so the binomial distribution with parameters n and p can be approximated by a Poisson distribution with parameter np. Note that as some number of X must be realized

$$e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1.$$

For the expectation and variance of X, we compute

$$\begin{aligned}\mathbf{E}[X] &= e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda, \\ \mathbf{E}[X(X-1)] &= e^{-\lambda} \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} = e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2, \\ \mathbf{Var}[X] &= \mathbf{E}[X(X-1)] + \mathbf{E}[X] - (\mathbf{E}[X])^2 = \lambda.\end{aligned}$$

For the Wright-Fisher model with constant population size we have $\lambda = np = 2N \cdot 1/2N = 1$. I.e. the average number of offspring is $\lambda = 1$, as it must be. The Possion distribution tells us that also the variance is $\lambda = 1$.

Here comes a less formal explanation for the offspring distribution: Let's first of all assume that the population is large with $2N$ individuals, $2N$ being larger than 30, say (otherwise offspring numbers will follow a binomial distribution and the above approximation to the Poisson does not work). Now all $2N$ individuals in generation $t+1$ will choose a parent among the individuals in generation t. We concentrate on one of the possible parents. The probability that a child chooses this parent is $\frac{1}{2N}$, and the probability that the child chooses a different parent is therefore $1 - \frac{1}{2N}$. The probability that also the second

child will not choose this parent is $(1-\frac{1}{2N})^2$. And the probability that all $2N$ children will not choose this parent is $(1-\frac{1}{2N})^{2N}$. And using the approximation from Maths 1.1 we can rewrite this (because as long as x is small, no matter if it is negative or positive, $1+x \approx e^x$) to $e^{\frac{-2N}{2N}} = e^{-1}$ (corresponding to the term $k=0$ of the Poisson distribution with parameter $\lambda = 1$).

A parent has exactly one offspring when one child chooses it as its parent and all other children do not choose it as the parent. Let us say that the first child chooses it as a parent, this has again probability $\frac{1}{2N}$. And also all the other individuals do not choose the parent, which then has probability $\frac{1}{2N} \cdot (1-\frac{1}{2N})^{2N-1}$. However, we should also take into account the possibility that another child chooses the parent and all others don't choose it. Therefore the probability that a parent has one offspring is $2N$ times the probability that only the first child chooses it: $2N \cdot (\frac{1}{2N}) \cdot (1-\frac{1}{2N})^{2N-1}$. This can be approximated as e^{-1} (the term corresponding to $k=1$ of the Poisson distribution).

The probability that a parent has 2 offspring which are child number 1 and child number 2 is $(\frac{1}{2N})^2 \cdot (1-\frac{1}{2N})^{2N-2}$ because for each of these children the probability of choosing the parent is $\frac{1}{2N}$ and all others should not choose this parent. In order to get the probability that any two children belong to this parent we just have to multiply with the number of ways we can choose 2 children out of $2N$ individuals which is $\binom{2N}{2}$. So the probability of a parent having 2 offspring is $\binom{2N}{2}(\frac{1}{2N})^2 \cdot (1-\frac{1}{2N})^{2N-2} \approx \frac{1}{2}e^{-1}$ (the term corresponding to $k=2$ of the Poisson distribution). You can continue like this and find every term of the Poisson distribution. We will return to the Poisson distribution when we describe the number of mutations on a branch in a tree in section 2.4.

Exercise 2.1. Try out `wf.model()` from the `R` package `labpopgen` which comes with this course. Look at the helpfile of `wf.model` by saying `?wf.model` and type `q` to get out of the help mode again. To use the function with the standard parameters, just type `wf.model()`.

- Does the number of offspring really follow a Poisson distribution?

□

Exercise 2.2. The Wright-Fisher model as we introduced it here is a model for haploid populations. Assume we also want to model diploids in the model. Can you draw a similar figure as Figure 2.1 for the diploid model? How do you need to update rules 1.-3. for this model? □

2.2 Genetic Drift

Genetic drift is the process of random changes in allele frequencies in populations. It can make alleles fix in the population or disappear from it. Drift is a stochastic process, which means that even though we understand how it works, there is no way to predict what will happen in a population with a specific allele. It is important to understand what this means for evolutionary biology: even if we would know everything about a population, and we would have a perfect understanding of the laws of biology, we cannot predict the state

of the population in the future. In this subsection, we introduce drift in several different ways so that you will get a feeling for its effects and the time scale at which these effects work. To describe drift mathematically, we again work with the binomial distribution.

Suppose you are looking at a small population of population size $2N = 10$. Now, if in generation 1 the frequency of A is 0.5, then what is the probability of having 0, 1 or 5 A's in the next generation?

This probability is given by the binomial sampling formula (in which $2N$ is the population size and p the frequency of allele A and therefore the probability that an individual picks a parent with genotype A). Let us calculate the expectation and the variance of a binomial distribution.

Maths 2.2. *Recall the binomial distribution from Maths 1.5. For the expectation and the variance of a binomial distribution with parameters n and p we calculate*

$$k\binom{n}{k} = \frac{k \cdot n!}{k!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1},$$
$$k(k-1)\binom{n}{k} = \frac{n!}{(k-2)!(n-k)!} = n(n-1)\binom{n-2}{k-2}.$$

Using this, the expectation is calculated to be

$$\begin{aligned}\mathbf{E}[X] &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)} \\ &= np\sum_{k=0}^{n}\binom{n-1}{k}p^k(1-p)^{n-1-k} = np\end{aligned}$$

and for the variance

$$\begin{aligned}\mathbf{E}[X^2 - X] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} \\ &= n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{(n-2)-(k-2)} = n(n-1)p^2\end{aligned}$$

and so

$$\begin{aligned}\mathbf{V}[X] &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mathbf{E}[X^2 - X] + \mathbf{E}[X] - \mathbf{E}[X]^2 = n(n-1)p^2 + np - n^2p^2 \\ &= np - np^2 = np(1-p).\end{aligned}$$

When simulating allele frequencies in a Wright-Fisher population, we don't need to pick a random parent for each individual one by one. We can just pick a random number from the binomial distribution (with the appropriate $2N$ and p) and use this as the frequency of the allele in the next generation. (If there is more than two different alleles, we use the multinomial distribution.) The binomial distribution depends on the frequency of the

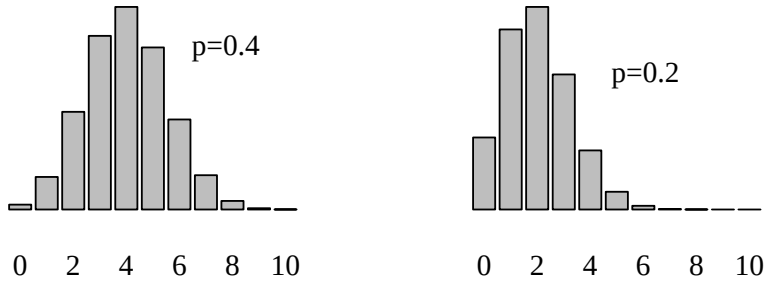


Figure 2.4: The binomial distribution for different parameters. Both have $n = 10$, the left one for $p = 0.2$ and the right one for $p = 0.5$.

allele in the last generation which enters as p, and on the population size which enters as $2N$. Obviously, for the special case $p = 1/2N$, we just get the offspring distribution of a single individual. Figure 2.4 shows two plots of the binomial distribution. As you can see, the probability of loosing the allele is much higher if p is smaller.

Exercise 2.3. Use `wf.model()` from the `R-` package to simulate a Wright Fisher population. You can change the number of individuals and the number of generations.

1. Pick one run, and use `untangeld=TRUE` to get the untangled version. Now suppose that in the first generation half of your population carried an A allele, and the other half an a allele. How many A-alleles do you then have in the 2nd, 3rd etc generation? It is easy to follow the border between the two alleles in the population. If you draw the border with a pencil you see that it is moving from left to right (and from right to left).

2. Try out different population sizes in `wf.model()`. Do the changes in frequency get smaller or bigger when you increase or decrease population size? Why is this the case? And how does the size of the changes depend on the frequency?

3. Consider an allele with frequency 0.1 in a population of size 10. What is the probability that the allele is lost in one generation? Assume the population size is 1000. What is the probability of loss of the allele now?

□

The random change of allele frequencies in a population from one generation to another is called *genetic drift*. Note that in the plots made by `wf.model()`, time is on the vertical

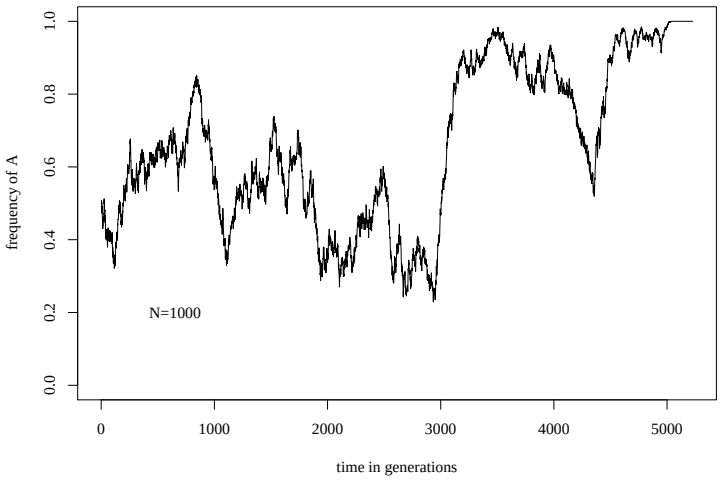


Figure 2.5: Frequency curve of one allele in a Wright-Fisher Model. Population size is $2N = 2000$ and time is given in generations. The initial frequency is 0.5.

axis whereas in the plots made by `wf.freq()`, time is on the horizontal axis. Usually if we plot frequencies that change in time, we will have the frequencies on the y-axis and time on the x-axis, so that the movement (drift) is vertical. Such a plot is given in Figure 2.5.

Exercise 2.4. What is your guess: Given an allele has frequency 0.5 in a population what is the (expected) time until the allele is lost or fixed in a population of size $2N$ compared to a population with twice that size? To do simulations, use

```
>res=wf.freq(init.A =0.5, N = 50, stoptime = 500, batch = 100)
>plot(res, what=c( "fixed" ) )
```

□

2.3 The coalescent

Until now, in our explanation of the Wright-Fisher model, we have shown how to predict the state of the population in the next generation $(t + 1)$ given that we know the state in the last generation (t). This is the classical approach in population genetics that follows the evolutionary process forward in time. This view is most useful if we want to predict the evolutionary outcome under various scenarios of mutation, selection, population size and structure, etc. that enter as parameters into the model. However, these model parameters are not easily available in natural populations. Usually, we rather start out with data from a present-day population. In molecular population genetics, this will be mostly sequence polymorphism data from a population sample. The key question then becomes: What are the evolutionary forces that have shaped the observed patterns in our data? Since

these forces must have acted in the history of the population, this naturally leads to a genealogical view of evolution backward in time. This view in captured by the so-called coalescent process (or simply *the coalescent*), which has caused a small revolution in molecular population genetics since its introduction in the 1980's. There are three main reasons for this:

- The coalescent is a valuable mathematical tool to derive analytical results that can be directly linked to observed data.
- The coalescent leads to very efficient simulation procedures.
- Most importantly, the coalescent allows for an intuitive understanding of population genetic processes and the patterns in DNA polymorphism that result from these processes.

For all these reasons, we will introduce this modern backward view of evolution in parallel with the classical forward picture.

The coalescent process describes the genalogy of a population sample. The key event of this process is therefore that, going backward in time, two or more individuals share a common ancestor. We can ask, for example: what is the probability that two individuals from the population today (t) have the same ancestor in the previous generation ($t-1$)? For the neutral Wright-Fisher model, this can easily be calculated because all individuals pick a parent at random. If the population size is $2N$ the probability that two individuals choose the same parent is

$$p_{c,1} = \mathbf{P}[\text{common parent one generation ago}] = \frac{1}{2N}. \tag{2.1}$$

Given the first individual picks its parent, the probability that the second one picks the same one by chance is 1 out of $2N$ possible ones. This can be iterated into the past. Given that the two individuals did not find a common ancestor one generation ago maybe they found one two generations ago and so on. We say that the lines of descent from the two individuals *coalescence* in the generation where they find a common ancestor for the first time. The probability for coalescence of two lineages exactly t generations ago is therefore

$$p_{c,t} = \mathbf{P}\left[\begin{array}{l}\text{Two lineages coalesce}\\ t \text{ generations ago}\end{array}\right] = \frac{1}{2N}\cdot\underbrace{\left(1-\frac{1}{2N}\right)\cdot\ldots\cdot\left(1-\frac{1}{2N}\right)}_{t-1 \text{ times}}$$

Mathematically, we can describe the *coalescence time* as a random variable that is geometrically distributed with success probability $\frac{1}{2N}$.

Maths 2.3. *If a random variable X is geometrically distributed with parameter p then*

$$\mathbf{P}[X=t] = (1-p)^{t-1}p, \qquad \mathbf{P}[X>t] = (1-p)^t,$$

i.e. the geometrical distribution gives the time of the first success for the successive performance of an experiment with success probability p.

Figure 2.6 shows the common ancestry in the Wright-Fisher animator from `wf.model()` In this case the history of just two individuals is highlighted. Going back in time there is always a chance that they choose the same parent. In this case they do so after 11 generations. In all the generations that follow they will automatically also have the same ancestor. The common ancestor in the 11th generation in the past is therefore called the *most recent common ancestor* (MRCA).

Exercise 2.5. What is the probability that two lines in Figure 2.6 coalesce exactly 11 generations in the past? What is the probability that it takes at least 11 generations for them to coalesce? □

The coalescence perspective is not restricted to a sample of size 2 but can be applied for any number $n(\leq 2N)$ of individuals. We can construct the genealogical history of a sample in a two-step procedure:

1. First, fix the topology of the coalescent tree. I.e., decide (at random), which lines of descent from individuals in a sample coalesce first, second, etc., until the MRCA of the entire sample is found.

2. Second, specify the times in the past when these coalescence events have happened. I.e., draw a so-called coalescent time for each coalescent event. This is independent of the topology.

For the Wright-Fisher model with $n \ll 2N$, there is a very useful approximation for the construction of coalescent trees that follows the above steps. This approximation relys on the fact that we can ignore multiple coalescence events in a single generation and coalescence of more than two lineages simultaneously (so-called "multiple mergers"). It is easy to see that both events occur with probability $(1/2N)^2$, which is much smaller than the simple coalescence probability of two lines.

With only pairwise coalescence events, the topology is easy to model. Because of neutrality, all pairs of lines are equally likely to coalesce. As the process is iterated backward in time, coalescing lines are combined into equivalence classes. We obtain a random bifurcating tree. Each topology can be represented by an expression in nested parentheses. For example, in a sample of 4, the expression $(((1,2),3),4)$ indicates that backward in time first lines 1 and 2 coalesce before both coalesce with 3 and these with 4. In $((1,3)(2,4))$, on the other hand, first pairs $(1,3)$ and $(2,4)$ coalesce before both pairs find a common ancestor.

For the branch lengths of the coalescent tree, we need to know the coalescence times. For a sample of size n, we need $n-1$ times until we reach the MRCA. As stated above, these times are independent of the topology. Mathematically, we obtain these times most conveniently by an approximation of the geometrical distribution by the exponential distribution for large N.

Maths 2.4. *There is a close relationship between the geometrical and the exponential distribution (see Maths 2.3 and Maths 1.3). If X is geometrically distributed with small*

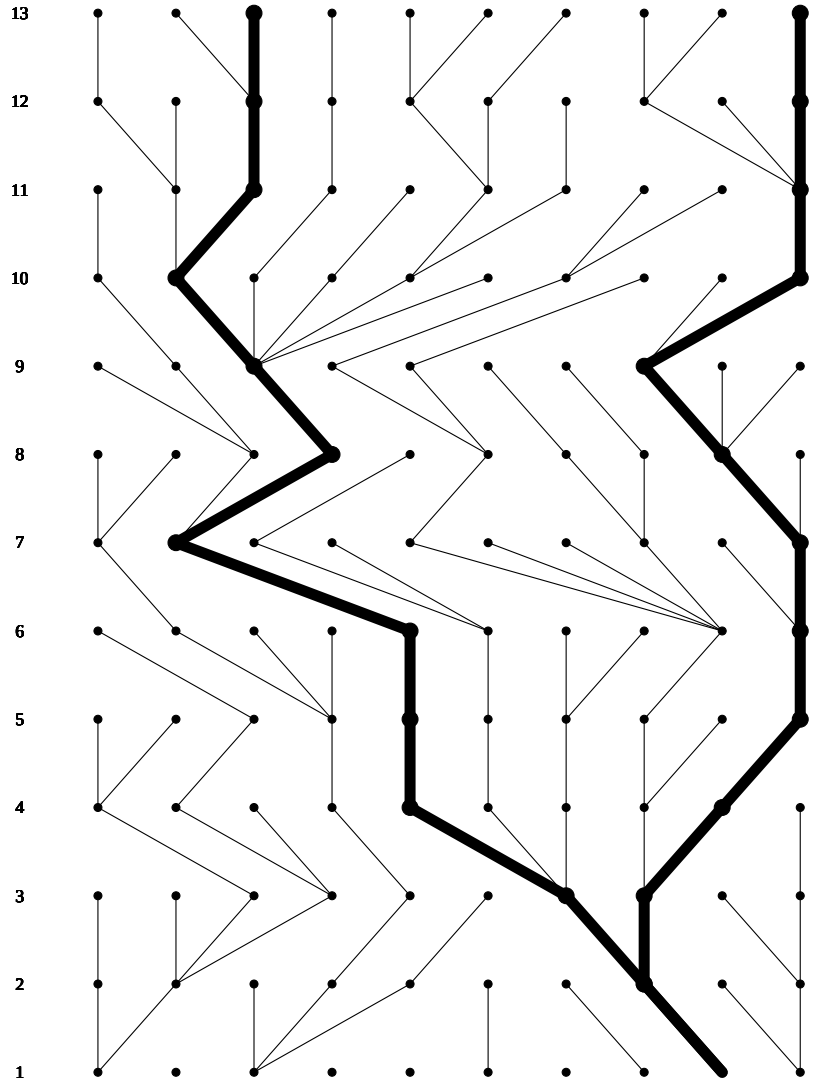


Figure 2.6: The coalescent of two lines in the Wright-Fisher Model

success probability p and t is large then

$$\mathbf{P}[X \geq t] = (1-p)^t \approx e^{-pt}.$$

This is the distribution function of an exponential distribution with parameter p.

For a sample of size n, there are $\binom{n}{2}$ possible coalescent pairs. The coalescent probability per generation is thus

$$\mathbf{P}[\text{coalescence in sample of size } n] = \frac{\binom{n}{2}}{2N}.$$

Let T_n be the time until the first coalescence occurs. Then

$$\mathbf{P}[T_n > t] = \left[1 - \frac{\binom{n}{2}}{2N}\right]^t \xrightarrow{N\to\infty} \exp\Big(-\frac{t\binom{n}{2}}{2N}\Big) \tag{2.2}$$

where we have used the approximation from Maths 1.1 which works if N is large. That means that in a sample of size n the waiting time until the first coalescence event is approximately exponentially distributed with rate $\frac{\binom{n}{2}}{2N}$. For the time from the first to the second coalescence event, T_{n-1}, we simply iterate this procedure with n replaced by $n-1$, etc.

Exercise 2.6. What is the coalescence rate for a sample of 6 (and population size $2N$)? What is the expected time you have to wait to go from 6 to 5 lineages? And from 5 to 4, 4 to 3, 3 to 2 and 2 to 1? Draw an expected coalescent tree for a sample of 6, using the expected waiting times for two different tree topologies. □

The tree in Figure 2.6 or the tree you have drawn in Exercise 2.6 is called a genealogical tree. A genealogical tree shows the relationship between two or more sequences. Don't confound it with a phylogenetic tree that shows the relationship between two or more species. The genealogical tree for a number of individuals may look different at different loci (whereas there is only one true phylogeny for a set of species). For example, at a mitochondrial locus your ancestor is certainly your mother and her mother. However, if you are a male, the ancestor for the loci on your Y-chromosome is your father and his father. So the genealogical tree will look different for a mitochondrial locus than for a Y-chromosomal locus. For a single locus, we are usually not able to reconstruct a single "true coalescence tree", but we can make inferences from the distribution of coalescence trees that are compatible with the data.

In order to get the tree in Figure 2.6, we did a forward in time simulation of a Wright-Fisher population and then extracted a genealogical tree for two individuals. This is very (computer) time consuming. By following the construction steps outlined above,it is also possible to do the simulation backward in time and only for the individuals in our sample. These coalescent simulations are typically much more efficient. Simulations in population genetics are important because they can be used to get the distribution of

certain quantities where we do not have the analytical results. These distributions in turn are used to determine whether data are in concordance with a model or not.

The fact that in the coalescent the times T_k are approximately exponentially distributed enables us to derive several important quantities. Below, we derive first the expected time to the MRCA and second the expected total tree length. The calculation uses results on the expectation and variance for exponentially distributed random variables from Maths 1.3.

Let T_k be the time to the next coalescence event when there are k lines present in the coalescent. Let further T_{MRCA} be the time to the MRCA and L the total tree length. Then

$$T_{MRCA} = \sum_{i=2}^{n} T_i, \qquad L = \sum_{i=2}^{n} iT_i. \tag{2.3}$$

So we can calculate for a coalescent of a sample of size n

$$\mathbf{E}[T_{MRCA}] = \sum_{i=2}^{n} \mathbf{E}[T_i] = \sum_{i=2}^{n} \frac{2N}{\binom{i}{2}} = \sum_{i=2}^{n} \frac{4N}{i(i-1)} = 4N \sum_{i=2}^{n} \frac{1}{i-1} - \frac{1}{i} = 4N\Big(1 - \frac{1}{n}\Big). \tag{2.4}$$

For the total tree length L we obtain

$$\mathbf{E}[L] = \sum_{i=2}^{n} i\mathbf{E}[T_i] = \sum_{i=2}^{n} i\frac{2N}{\binom{i}{2}} = 4N \sum_{i=2}^{n} \frac{1}{i-1} = 4N \sum_{i=1}^{n-1} \frac{1}{i}.$$

Note that even for a large sample

$$\mathbf{E}[T_{MRCA}] < 4N, \qquad \mathbf{E}[T_2] = 2N,$$

so that in expectation more than half of the total time in the coalescent till the MRCA is needed for two remaining ancestral lines to coalesce. Also the variance in T_{MRCA} is dominated by the variance in T_2. For larger samples, the expected time to the MRCA quickly reaches a limit. A related result is that the probability that the coalescent of a sample of size n contains the MRCA of the whole population is $(n-1)/(n+1)$ (for large, finite N). Increasing the sample size will mostly add short twigs to a coalescent tree. As a consequence, also the total branch length

$$\mathbf{E}[L] \approx 4N \log(n-1).$$

increases only very slowly with the sample size. An important practical consequence of these findings is that, under neutrality, relatively small sample sizes (typically 10-20) will usually be enough to gain all statistical power that is available from a single locus.

Exercise 2.7. The true coalescent tree doesn't have to look like the expected tree. In fact it is unlikely that any random tree looks even similar to the expected tree. Use `coalator()` from the `R`-package to simulate a couple of random trees for a sample of 6 sequences. Produce 10 trees with sample size 6. Write down for every tree that you simulate its depth (i.e. the length from the root to a leaf). How much larger approximately is the deepest tree compared to the shallowest tree you made? Do the same comparison for the time in the tree when there are only 2 lines. □

Exercise 2.8. The variance is a measure of how variable a random quantity, e.g., the depth of a coalescent tree, is. Two rules, which are important to compute variances, are for independent random quantities X and Y,

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y], \qquad \mathbf{Var}[iX] = i^2\mathbf{Var}[X].$$

The depth is the same as the time to the MRCA, so consider T_{MRCA} as given in (2.3). Can you calculate the variance of the two quantities you measured in the last exercise? □

2.4 Mutations in the infinite sites model

When we described the Wright-Fisher Model, we left out mutation all together. We can easily account for neutral mutations, however, by simply changing the update rule to

3'. With probability $1-\mu$ an offspring takes the genetic information of the parent. With probability μ it will change its genotype.

This rule is unspecific in how the change looks like. In the simplest case, we imagine that an individual can only have one of two states, for example a and A, which could represent wildtype and mutant. Depending on the data we deal with, we can choose a model that tells us which changes are possible. The standard model for DNA sequence data is the *infinite sites model*. The key assumption of the infinite sites model is that every new mutation hits a new site in the genome. It therefore cannot be masked by recurrent or back-mutations and will be visible in the population unless it is lost by drift. Whether the infinite site assumption is fulfilled depends on the mutation rate and the evolutionary time scale we are concerned with.

Let us now see how mutations according to the infinite sites scheme can be introduced in the coalescent framework. It is useful to define a mutation rate that is scaled by the population size. In the following exercise, we consider first a single line of descent:

Exercise 2.9. Follow back one line in the coalescent. Assuming mutations occur with probability μ per generation what is the probability that the line is not hit by a mutation by time t? Can you approximate this probability? What is the distribution of the waiting time to the first mutation event? □

Assume now that we have a coalescent tree of a sample of size n. In order to get a sample with polymorphic sites, we want to add mutations to this tree. For any given branch of the tree we could do this by repeatedly drawing random numbers for the waiting time to a mutation from an exponential distribution and adding mutations as long as the branch length exceeds the cumulated waiting time. The mutation will be visible in all descendents from that branch. The crucial point is that, for *neutral* mutations, we can do this without interfering with the shape or size of the tree (i.e. its topology and the branch lengths). The reason is that, forward in time, a neutral mutation does not change the offspring distribution of an individual. Consequently, it does not change its probability to be picked as a parent backward in time. Under neutrality, *state* (the genotype of an individual)

and *descent* (the genealogical relationships) are independent stochastic processes. In the construction of a coalescent with mutations, they can be dealt with in seperate steps.

Usually, one is not so much interested in the exact times of mutation events, but rather in the number of mutations on each branch of the tree. We can make use of a close connection between the exponential and the Poisson distribution to address this quantity directly:

Maths 2.5. *Consider a long line starting at 0. After an exponential time with parameter λ a mark hits the line. After another time with the same distribution the same happens etc. Then the distribution of marks in an interval $[0,t]$ is Poisson distributed with parameter λt.*

For a branch of length l, we therefore directly get the number of neutral mutations on this branch by drawing a Poisson distributed random number with parameter $l\mu$. In particular, the total number of mutations in an entire coalescent tree of length L the tree Poisson distributed with parameter $L\mu$. Let S be the number of mutations on the tree. Then

$$\mathbf{P}[S=k] = \int_0^\infty \mathbf{P}[S=k|L\in d\ell]\mathbf{P}[L\in d\ell] = \int_0^\infty e^{-\ell\mu}\frac{(\ell\mu)^k}{k!}\mathbf{P}[L\in d\ell].$$

For the expectation that means

$$\begin{aligned}\mathbf{E}[S] &= \sum_{k=0}^\infty k\mathbf{P}[S=k] = \int_0^\infty \ell\mu e^{-\ell\mu}\Big(\sum_{k=1}^\infty \frac{(\ell\mu)^{k-1}}{(k-1)!}\Big)\mathbf{P}[L\in d\ell] \\ &= \mu\int_0^\infty \ell\ \mathbf{P}[L\in d\ell] = \frac{\theta}{4N}\mathbf{E}[L] = \theta\sum_{i=1}^{n-1}\frac{1}{i}\end{aligned} \tag{2.5}$$

where

$$\theta = 4N\mu$$

is the standard population mutation parameter.

Estimators for the mutation rate

All population genetic models, whether forward or backward in time, depend on a set of biological parameters that must be estimated from data. In our models so far, the two key parameters are the mutation rate and the population size. Both combine in the population mutation parameter θ. With the above equations at hand we can already define two estimators of θ.

Since the infinite sites model assumes that each mutation on the genealogical tree gives one new segregating site in the sample of DNA sequences. We can then estimate the parameter θ from the observed segregating sites in a sample using (2.5). Consider first a subsample of size 2 from our sample. For each such subsample we have

$$\mathbf{E}[S] = \theta.$$

Denote by S_{ij} the number of differences between sequence i and j. Since there are $\binom{n}{2}$ subsamples of size 2 in a sample of size n, we can define

$$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}} \sum_{i<j} S_{ij}. \tag{2.6}$$

$\widehat{\theta}_\pi$ is an unbiased estimator of θ based on the expected number of pairwise differences which is usually referred to as π. (In the literature, also the estimator is often called π, but we prefer to distinguish parameters and estimators here.)

Another unbiased estimator for θ can be read directly from (2.5):

$$\widehat{\theta}_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}. \tag{2.7}$$

This estimator was first described by WATTERSON (1975) using diffusion theory. Its origin becomes only apparent in the coalescent framework, however.

Exercise 2.10. Can you explain why the above estimators are unbiased? □

Exercise 2.11. Open the file `055.nex` with `DNASP`. You see sequences of 24 individuals of *Drosophila*, the first one from a line of *Drosophila Simulans*, 11 European lines from *Drosophila Melanogaster* and 12 from the African population of *Drosophila Melanogaster*. Compute $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ for the African and European subsamples (alternatively you can click on `Overview->Interspecific Data`, the estimates are displayed). The estimator $\widehat{\theta}_\pi$ is denoted `pi` and the estimator $\widehat{\theta}_S$ is denoted `Theta_W` (where W stands for its discoverer *Watterson*).

1. Look at the data. Can you also calculate $\widehat{\theta}_S$ by hand? And what about $\widehat{\theta}_\pi$? Which computation steps do you have to do here?

2. Instead of taking only the African subsample you can also take all 24 sequences and see what $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ is. Here you see that not the number of segregating sites S are used for computation of $\widehat{\theta}_S$ but the total number of mutations (which is called `eta` here). Why do you think that makes sense? Which model assumptions of the infinite site model are not met by the data.

3. What do you think about the estimators you get for the whole dataset? Do you expect these estimators to be unbiased?

4. The estimators for θ are much larger for the African population than for the European one. Can you think of an explanation for this?

□

3 Effective population size

In the first two chapters we have dealt with idealized populations. The two main assumptions were that the population has a constant size and the population mates panmictically. These ideal populations are good to start with because they allow us to derive some important results. However, natural populations are usually not panmictic and the population size may not be constant over time. Nevertheless, we can often still use the theory that we have developed. The trick is that we describe a natural population as if it is an ideal one by adjusting some parameters, in this case the population size. This is the idea of the effective population size which is the topic of this section.

Human Population Size Example

As an example, we will analyse a dataset from HAMMER *et al.* (2004). The dataset, which may be found in the file `TNFSF5.nex`, contains data from different human populations: Africans, Europeans, Asians and Native Americans. It contains 41 sequences from 41 males, from one locus, the TNFSF5 locus. TNFSF5 is a gene and the sequences are from the introns of this gene, so there should be no constraint on these sequences, in other words every mutation should be neutral. The gene lies on the X-chromosome in a region with high recombination. What that means for the data will become clearer later in the course.

Exercise 3.1. Import the data in `DNASP` and determine $\widehat{\theta}_\pi$ per site and $\widehat{\theta}_S$ per site using all 41 sequences. □

As you have seen in Section 2, both $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ are estimators of the model parameter $\theta = 4N\mu$ where μ is the probability that a site mutates in one generation. However, the TNFSF5 locus is on the X-chromosome and for the X-chromosome males are haploid. Therefore the population of X-chromosomes can be seen as a population of $1.5N$ haploids (instead of $2N$ haploids for autosomes) and therefore in this case $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ are estimators of $3N\mu$. The reason that HAMMER *et al.* (2004) looked at X-chromosomes is mainly because the sequencing is relatively easy. Males have only one X-chromosome, so you don't have to worry about polymorphism within one individual (more about polymorphism within an individual in Section 4).

The mutations in these data are single nucleotide polymorphisms. SNPs are frequently used to determine $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ per site. Their (per site) mutation rate is estimated to be $\mu = 2 \cdot 10^{-8}$ by comparing human and chimpanzee sequences.

Exercise 3.2. Recall Section 1. Assume that the divergence time of chimpanzees and humans is 10MY with a generation time of 20 years and the mutation rate is $2 \cdot 10^{-8}$ per nucleotide per generation..

1. What is the expected percentage of sites that are different (or what is the divergence) between human and chimp?

2. Both $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ are estimators for $3N\mu$ and both can be directly computed from the data. What estimate of N do you get, when using the estimated $\widehat{\theta}$ values from the last exercise?

3. There are about $6 \cdot 10^9$ people on earth. Does the human population mate panmictically? Is the population constant over time? Can you explain why your estimate of N is so different from $6 \cdot 10^9$?

□

The number of $6 \cdot 10^9$ people on earth is referred to as the *census population size*. This section is about a different notion of population size which is called the *effective* population size.

3.1 The concept

Before we start with calculations using *effective* population sizes we introduce what they are. We use the following philosophy:

> Let • be some measurable quantity that relates to the strength of genetic drift in a population. This can be e.g. the rate of loss of heterozygosity or the probability of identity by descent. Assume that this quantity has been measured in a natural population. Then the effective size N_e of this population is the size of an ideal (neutral panmictic constant-size equilibrium) Wright-Fisher population that gives rise to the same value of the measured quantity •. To be specific, we call N_e the •-effective population size.

In other words, the effective size of a natural population is the size of the ideal population such that some key measure of genetic drift is identical. With an appropriate choice of this measure we can then use a model based on the ideal population to make predictions about the natural one. Although a large number of different concepts for an effective population size exist, there are two that are most widely used.

The identity-by-descent (or inbreeding) effective population size

One of the most basic consequences of a finite population size - and thus of genetic drift - is that there is a finite probability for two randomly picked individuals in the offspring generation to have a common ancestor in the parent generation. This is the *probability of identity by descent*, which translates into the single-generation coalescence probability of two lines $p_{c,1}$ in the context of the coalescent. For the ideal Wright-Fisher model with $2N$ (haploid) individuals, we have $p_{c,1} = 1/2N$. Knowing $p_{c,1}$ in a natural population, we can thus define the identity-by-descent effective population size

$$N_e^{(i)} = \frac{1}{2p_{c,1}}. \tag{3.1}$$

We will see in the next chapter that the degree of inbreeding is one of the factors that influences $N_e^{(i)}$. For historic reasons, $N_e^{(i)}$ is therefore usually referred to as *inbreeding effective population size*. Since all coalescent times are proportional to the inverse coalescent probability, they are directly proportional to the inbreeding effective size. One also says that $N_e^{(i)}$ fixes the *coalescent time scale*.

The variance effective population size

Another key aspect about genetic drift is that it leads to random variations in the allele frequencies among generations. Assume that p is the frequency of an allele A in an ideal Wright-Fisher population of size $2N$. In Section 2, we have seen that the number of A alleles in the next generation, $2Np'$, is binomially distributed with parameters $2N$ and p, and therefore

$$\mathbf{Var}_{WF}[p'] = \frac{1}{(2N)^2}\mathbf{Var}[2Np'] = \frac{p(1-p)}{2N}.$$

For a natural population where the variance in allele frequencies among generations is known, we can therefore define the variance effective population size as follows

$$N_e^{(v)} = \frac{p(1-p)}{2\mathbf{Var}[p']}. \tag{3.2}$$

As we will see below, the inbreeding and variance effective sizes are often identical or at least very similar. However, there are exceptions and then the correct choice of an effective size depends on the context and the questions asked. Finally, there are also scenarios (e.g. changes in population size over large time scales) where no type of effective size is satisfactory. We then need to abandon the most simple ideal models and take these complications explicitly into account.

Loss of heterozygosity

As an application of the effective-population-size concept, let us study the loss of heterozygosity in a population. Heterozygosity H can be defined as the probability that two alleles, taken at random from a population are different at a random site (or locus). Suppose that the heterozygosity in a natural population in generation 0 is H_0. We can ask, what is the expected heterozygosity in generation $t = 1, 2, 3$, if we assume no new mutation (i.e. we only consider the variation that is already present in generation 0). In particular, for $t = 1$, we find

$$H_1 = \frac{1}{2N_e^{(i)}}0 + \Big(1 - \frac{1}{2N_e^{(i)}}\Big)H_0 = \big(1 - \frac{1}{2N_e^{(i)}}\big)H_0. \tag{3.3}$$

Indeed, if we take two random alleles from the population in generation 1, the probability that they have the same parent in generation 0 is $\frac{1}{2N_e^{(i)}}$. When this is the case they have probability 0 to be different at any site because we ignore new mutations. With probability

$1 - \frac{1}{2N_e^{(i)}}$ they have different parents in generation 0 and these parents have (by definition) probability H_0 to be different. By iterating this argument, we obtain

$$H_t = \Big(1 - \frac{1}{2N_e^{(i)}}\Big)^t \cdot H_0$$

for the heterozygosity at time t. This means that, in the absence of mutation, heterozygosity is lost at rate $\frac{1}{2N_e^{(i)}}$ every generation and depends only on the inbreeding effective population size.

Estimating the effective population size

For the Wright-Fisher model, we have seen in Section 2 that the expected number of segregating sites S in a sample is proportional to the mutation rate and the total expected length of the coalescent tree, $\mathbf{E}[S] = \mu[L]$. The tree-length L, in turn, is a simple function of the coalescent times, and thus of the inbreeding effective population size $N_e^{(i)}$. Under the assumption of (1) the infinite sites model (no double hits), (2) a constant $N_e^{(i)}$ over the generations (constant coalescent probability), and (3) a homogeneous population (equal calescent probability for all pairs) we can therefore estimate the effective population size from polymorphism data if we have independent knowledge about the mutation rate (e.g. from divergence data). In particular, for a sample of size 2, we have $\mathbf{E}[S_2] = 4N_e^{(i)}\mu$ and thus

$$N_e^{(i)} = \frac{\mathbf{E}[S_2]}{4\mu}.$$

In a sample of size n, we can estimate the expected number of pairwise differences to be $\widehat{\mathbf{E}}[S_2] = \widehat{\theta}_\pi$ (see (2.6)) and obtain the estimator of $N_e^{(i)}$ from polymorphism data as

$$\widehat{N}_e^{(i)} = \frac{\widehat{\theta}_\pi}{4\mu}.$$

A similar estimate can be obtained from Watterson's estimator $\widehat{\theta}_S$, see Eq. (2.7). While the assumption of the infinite sites model is often justified (as long as $4N_e^{(i)}\mu_n \ll 1$, with μ_n the per-nucleotide mutation rate), the assumption of constant and homogeneous coalescent rates is more problematic. We will come back to this point in the next section when we discuss variable population sizes and population structure.

3.2 Examples

Let us now discuss the main factors that influence the effective population size. For simplicity, we will focus on $N_e^{(i)}$. We will always assume that there is only a single deviation from the ideal Wright-Fisher population.

Offspring variance

One assumption of the ideal model is that the offspring distribution for each individual is Binomial (approximately Poisson). In natural populations, this will usually not be the case. Note that average number of offspring must always be 1, as long as we keep the (census) population size constant. The offspring variance σ^2, however, can take any value in a wide range. Let X_i be the number of offspring of individual i with $\sum_i m_i = 2N$. Then the probability that individual i is the parent of two randomly drawn individuals from the offspring generation is $m_i(m_i-1)/(2N(2N-1))$

$$\sum_{i=1}^{2N} \frac{m_i(m_i-1)}{2N(2N-1)} \tag{3.4}$$

is the probability for identity by descent of two random offspring. The single-generation coalescent probability $p_{c,1}$ is the expectation of this quantity. With $\mathbf{E}[m_i] = 1$ and $\mathbf{E}[m_i^2] = \sigma^2 + 1$ and the definition (3.1) we arrive at

$$N_e^{(i)} = \frac{N - 1/2}{\sigma^2} \approx \frac{N}{\sigma^2}. \tag{3.5}$$

By a slightly more complicated derivation (not shown), we can establish that the variance effective population size $N_e^{(v)}$ takes the same value in this case.

Separate sexes

A large variance in the offspring number leads to consequence that in any single generation some individuals contribute much more to the offspring generation than others. So far, we have assumed that the offspring distribution for all individuals is identical. In particular, the expected contribution of each individual to the offspring generation was equal (= 1). Even without selection, this is not necessarily the case. An important example are populations with seperate sexes and unequal sex ratios in the breeding population. Consider the following example:

Imagine a zoo population of primates with 20 males and 20 females. Due to dominance hierarchy only one of the males actually breeds. What is the inbreeding population size that informs us, for example, about loss of heterozygosity in this population? 40? or 21??

Let N_f be the number of breeding females (20 in our case) and N_m the number of breeding males (1 in the example). Then half of the genes in the offspring generation will derive from the N_f parent females and half from the N_m parent males. Now draw two genes at random from two individuals of the offspring generation. The chance that they are both inherited from males is $\frac{1}{4}$. In this case, the probability that they are copies from the same paternal gene is $\frac{1}{2N_m}$. Similarly, the probability that two random genes are descendents from the same maternal gene is $\frac{1}{4}\frac{1}{2N_f}$. We thus obtain the probability of finding a common ancestor one generation ago

$$p_{c,1} = \frac{1}{4}\frac{1}{2N_m} + \frac{1}{4}\frac{1}{2N_f} = \frac{1}{8}\Big(\frac{1}{N_m} + \frac{1}{N_f}\Big)$$

and an effective population size of

$$N_e^{(i)} = \frac{1}{2p_{c,1}} = \frac{4}{\frac{1}{N_m} + \frac{1}{N_f}} = \frac{4N_f N_m}{N_f + N_m}.$$

In our example with 20 breeding females and 1 breeding male we obtain

$$N_e^{(i)} = \frac{4 \cdot 20 \cdot 1}{20 + 1} = \frac{80}{21} \approx 3.8.$$

The identity-by-decent (or inbreeding) effective population size is thus much smaller than the census size of 40 due to the fact that all offspring have the same father. Genetic variation will rapidly disappear from such a population. In contrast, for an equal sex ratio of $N_f = N_m = \frac{N}{2}$ we find $N_e = N$.

Sex chromosomes and organelles

Take two random Y-chromosome alleles from a population. What is the probability that they have the same ancestor one generation ago? This is the probability that they have the same father, because Y-chromosomes come only from the father. So this probability is $\frac{1}{N_m}$ where N_m is the number of males in the population, so $N_e^{(i)} = N_m$. Similarly, for mitochondrial genes $N_e^{(i)} = N_f$ where N_f is the number of females in the population. In birds the W-chromosome is the female determining chromosome. WZ individuals are female and ZZ individuals are male. So for the W-chromosome $N_e = N_f$. For the X-chromosome in mammals and the Z-chromosome in birds it is a bit more complicated. Take two random X-chromosome alleles, what is the probability that they are from the same ancestor? This is

$$\frac{1}{2}\Big(\frac{N_f}{N_m} + \frac{N_m}{N_f}\Big) \cdot \frac{1}{2N_f + N_m}.$$

Exercise 3.3. Explain the last formula. (Hint: You need to treat males and females in both the offspring and parent generations seperately.) What is $N_e^{(i)}$ for the X-chromosome if the sex ratio is 1:1? □

Fluctuating Population Sizes

Another assumption of the ideal Wright-Fisher model is a constant population size. Of course, the population size of most natural populations will rather fluctuate over time. In this case, the census size - and also the effective size - of a population changes from generation to generation. However, if fluctuations in the population size occur over cycles of only a few generations, it makes sense to define a single *long-term* effective population size to capture the average effect of drift over longer evolutionary periods.

Consider the evolution of a population with varying size over t generations and imagine that we have already calculated the effective population size for each individual generation

N_0 to N_{t-1}. The N_i take breeding structure etc. into account. The expected reduction of heterozygosity over t generations then is

$$\begin{aligned} H_t &= \Big(1 - \frac{1}{2N_0}\Big) \cdots \Big(1 - \frac{1}{2N_{t-1}}\Big) H_0 \\ &= (1 - \bar{p}_{c,1})^t H_0 \end{aligned}$$

where $\bar{p}_{c,1}$ is the relevant average single-generation coalescence probability that describes the loss of heterozygosity. We then have

$$\begin{aligned} 1 - \bar{p}_{c,1} &= \Big[\Big(1 - \frac{1}{2N_0}\Big) \cdots \Big(1 - \frac{1}{2N_{t-1}}\Big)\Big]^{1/t} \approx \Big[\exp\Big(-\frac{1}{2N_0}\Big) \cdots \exp\Big(-\frac{1}{2N_{t-1}}\Big)\Big]^{1/t} \\ &= \exp\Big(-\tfrac{1}{2t}\Big(\frac{1}{N_0} + \ldots + \frac{1}{N_{t-1}}\Big)\Big) \approx 1 - \frac{1}{2t}\Big(\frac{1}{N_0} + \ldots + \frac{1}{N_{t-1}}\Big) \end{aligned}$$

and get an average (inbreeding) effective population size of

$$\bar{N}_e^{(i)} = \frac{1}{2}\frac{1}{\bar{p}_{c,1}} \approx \frac{t}{\frac{1}{N_0} + \ldots + \frac{1}{N_{t-1}}}.$$

So in this case the (average) inbreeding-effective population size is given by the harmonic mean of the population sizes over time. Other than the usual arithmetic mean, the harmonic mean is most strongly influenced by single small values. E.g., if the N_i are given by 100, 4, 100, 100, the arithmetic mean is 76, but we obtain a harmonic mean of just $\bar{N}_e^{(i)} = 14$.

We can summarize our findings by the remark that many populations are genetically smaller than they appear from their census size, increasing the effects of drift.

Exercise 3.4. In Exercise 3.2 you estimated N_e for the X-chromosome in humans. The human population was not of constant size in the past. Assume that within the last 200000 years (i.e. within the last 10000 generations) the human population grew geometrically. That means that

$$N_{t+1} = gN_t.$$

How large must g be in order to explain your effective population size? □

Two toy models

Let us deal with two examples of populations that are unrealistic but helpful to understand the concept of effective sizes. The first example that is given represents a zoo population. In order to keep polymorphism as high as possible, care is taken that every parent has exactly one offspring. The ancestry is given by Figure 3.1(A).

The second example is similar to the example of unequal sex ratios where the population had 20 males, but only one of them had offspring. However, the next example is even more extreme. In this case the individuals are haploids, and only one ancestor is the parent of the whole next generation. A scenario like this is shown in Figure 3.1(B).

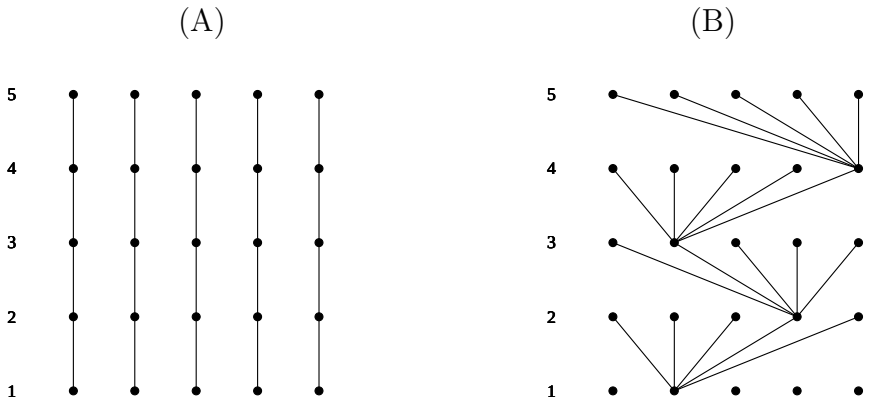


Figure 3.1: (A) The ancestry of a population that is kept as polymorphic as possible. (B) The ancestry of a population where each generation only has one parent

Exercise 3.5. Figures 3.1(A) and (B) clearly do not come from ideal Wright-Fisher populations. So $N_c \neq N_e$ and we can try to calculate the effective population sizes for them. Given the census size in the two cases is N_c what are the

- variance effective size,
- inbreeding effective size.

□

3.3 Effects of population size on polymorphism

We have already seen that genetic drift that removes variation from the population. So far, we have neglected mutation that creates new variation. In a natural population that evolves only under mutation and drift, an equilibrium between these two processes will eventually be reached. This equilibrium is called *mutation-drift balance*.

Mutation-drift balance

The neutral theory of molecular evolution tries to explain observed patterns in nucleotide frequencies across populations with only two main evolutionary forces: mutation and drift. Mutation introduces new variation, and drift causes them to spread, but also causes them to be lost. An equilibrium is found when these two processes balance. We have already derived that genetic drift reduces the heterozygosity each generation by $\Delta H = \frac{1}{2N_e^{(i)}} H$. In order to derive the equilibrium frequency, we still need to know the change in heterozygosity H under mutation alone. In the infinite sites mutation model, every new mutation hits a new site. Therefore, mutation can never reduce heterozygosity by making alleles identical.

However, every pair of identical alleles has the chance to become heterozygote if either of the genes mutates. Then, heterozygosity increases due to mutation like $\mathbf{E}[H'|H] = H + 2\mu(1-H)$. Summing over the effects of mutation and drift and ignoring terms of order μ^2 we obtain:

$$\mathbf{E}[H'|H] = H - \frac{1}{2N_e^{(i)}}H + 2\mu(1-H), \qquad \mathbf{E}[\Delta H|H] = -\frac{1}{2N_e^{(i)}}H + 2\mu(1-H).$$

The equilibrium is obtained for $\mathbf{E}[\Delta H|H] = 0$ and so

$$2\mu(1-H) - \frac{H}{2N_e^{(i)}} = 0, \qquad H\Big(2\mu + \frac{1}{2N_e^{(i)}}\Big) = 2\mu$$

and so the equilibrium heterozygosity is at

$$H^* = \frac{2\mu}{\Big(2\mu + \frac{1}{2N_e^{(i)}}\Big)} = \frac{\theta}{\theta+1}. \tag{3.6}$$

As expected, the equilibrium heterozygosity increases with increasing mutation rate (because more mutations enter the population) and with increasing effective population size (because of the reduced effect of drift). Note that only the product θ of both quantities enters the result.

There is an alternative way of deriving the same formula using the coalescent. What we are looking for is the probability that two individuals are different at some gene or at some nucleotide. If you follow their history back in time, two things can happen first: (1) either one of the two mutates or (2) they coalesce. If they coalesce first they are identical, if one of the two mutates they are not identical. Since mutation (in either lineage) occurs at rate 2μ and coalescence occurs at rate $1/(2N_e^{(i)})$, it is intuitive that the relative probabilities of both events to occur first are $2\mu : [1/(2N_e^{(i)})]$ resulting in the probability for mutation first as given in Eq. (3.6). For a rigorous mathematical treatment, we need the following result about exponential distributions:

Maths 3.1. *Let X and Y be exponentially distributed with rates μ and ν then*

$$\begin{aligned}\mathbf{P}[X < Y] &= \int_0^\infty \mathbf{P}[X = x] \cdot \mathbf{P}[Y \geq x]dx = \int_0^\infty \mu e^{-\mu x} e^{-\nu x} dx = \frac{\mu}{\mu+\nu} e^{-x(\mu+\nu)}\Big|_0^\infty \\ &= \frac{\mu}{\mu+\nu}.\end{aligned}$$

Assume X and Y are waiting times for certain events. Then the probability that X occurs before Y is determined by the fraction of the rate for X with respect to the total rate for both events.

Exercise 3.6. Using Maths 3.1 can you rederive (3.6)? □

3.4 Fixation probability and time

The probability that a new mutation that has occured in a population is not quickly lost again, but reaches fixation, and the time that it takes to do so, are fundamental quantities of molecular population genetics. Below, we will introduce these quantities for the neutral model.

Fixation probability of a neutral mutation

What is the fixation probability of a single new mutation in a population of (haploid) size $2N$ under neutrality? We can find the answer to this question by a simple argument that is inspired by genealogical thinking. Obviously, the mutation will eventually either fix in the population or get lost. Assume now that we move fast forward in time to a generation where this fate has certainly been sorted out. Now imagine that we draw the genealogical (or coalescent) tree for the entire population for this later generation. This genealogy will trace back to a single ancestor in the generation where the mutation that we are concerned with happened. Fixation of the mutation has occured if and only if the mutant is that ancestor. Since under neutrality each individual has the same chance to be picked as the ancestor, the neutral fixation probability must be

$$p_{\text{fix}} = \frac{1}{2N}.$$

You can have a look again at the simulation of `wf.model()` that you saw in chapter 2, and check how often the lower left individual is ancestor to all individuals in the last generation of the simulation.

Exercise 3.7. Use a similar argument to derive the fixation probability of a mutation that initially segregates at frequency p in the population.

Exercise 3.8. This exercise is about the equilibrium level of heterozygosity (or the maintenance of heterozygosity in the face of drift. It has always been (and still is) a major question in evolutionary biology why there is variation. Selection and drift are usually thought to remove variation, so when looking at data, the obvious question is always: how is variation maintained. You will be dealing with the theory developed above. For this exercise you will use the function `maint()` from the `R`-package.

If you type `maint()` , the function `maint()` is carried out with standard values $N = 100$, $u = 0.001$, `stoptime`=10, `init.A`=1, `init.C`=0, `init.G`=0, `init.T`=0 is simulated. You can e.g. change the popualtion size by using `maint(N=1000)`. If you want to plot the frequencies of the alleles over time, you can put the result of `maint()` directly into a plot command, for example by

```
>plot(maint(N=1000, stoptime=10000, u=0.00001))
```

If you just want to look at how the frequency of `G` changes, you can use the option `plot(maint(...), what="G")`.

This model simulates a population undergoing drift and continual mutation at a neutral locus. The locus is really just one nucleotide and it can therefore be in four states: `A`, `C`, `G` or `T`. At generation 0, all members of the population have the same homozygous genotype (i.e., the frequency of the single allele in the population is equal to one). In subsequent generations, new alleles enter the population through mutation.

1. Set the model parameters for a population of 300 individuals with a mutation rate at this locus of 10^{-4}. Observe the population for 10000 generations. View the graph of the frequencies of the different alleles in the population over time. Run this simulation a few times.

2. What happens to most new mutants that enter the population? How many alleles in this population attained frequencies above 0.1? Do any new mutant alleles reach a frequency above 0.9?

3. Based on this population size and mutation rate, what is the rate at which new mutants enter the population? (Note the appropriate formula as well as the numerical answer.) What is the rate at which you would expect new mutants to become fixed? You can also view the number of new mutations that occurred in the population (using `what="numOfMut"`). Does it fit with your expectation? Based on this rate, how many new mutants would you expect to become fixed during the 10000 generations for which you observed the population (check using `what="fixed"`)? Explain what this value means.

4. How does the number of fixed mutations depend on the population size? What does this mean for divergence between two species with different population sizes?

5. Now view the graph of heterozygosity over time (`what="h"`). What does this graph suggest about the level of variation in the population (i.e. is it fairly constant through time or does it change, is H above zero most of the time)? Give a rough estimate for the average value of H throughout these 10000 generations.

6. Using (3.6), predict the equilibrium value of H. You can also plot the predicted value using `what=c("h","h.pred")`.

7. Would you expect the heterozygosity to increase, decrease, or remain the same if you significantly increased the mutation rate? What would you expect if you increased the population size? What if you increase mutation rate but decrease population size?

8. Increase the mutation rate for the model to $5 \cdot 10^{-4}$. View the graph of heterozygosity over time and the graph of allele frequencies. Does this simulation confirm your expectation (given in the last task)? What does the formula predict H^* to be in this situation?

□

Fixation time under neutrality

In general, fixation times are not easily derived using the tools of this course. For the special case of a single neutral mutation in an ideal Wright-Fisher population one can show that the expected fixation time is equal to the expected time to the MRCA (the height of the coalescent tree) of the entire population. We have already calculated this quantity in section 2.3 (see eq. 2.4). If the sample size equals the population size $2N$, we need to take the large-sample limit and obtain

$$T_{\text{fix}} \approx 4N.$$

Exercise 3.9. Use `maint()` from the `R-` package. In this exercise you will look at the fixation time for a neutral mutation in a population. Set the mutation rate so that you a have only few (one or two) fixations in the graph. You need to set `stoptime` to a large value, maybe 100.000 and N to a low value, maybe 100.

1. Use $N = 100$ and look at least at 10 fixation events. Record the time it takes between appearance of the new muation and its fixation.

2. Take a different value for N and repeat the exercise. Plot (roughly) the relationship between N and the mean time to fixation for a neutral mutation.

   HINT: First create one instance of the evolution of the neutral locus using the command `res<-maint(...)`. Afterwards you can look at specific times during the evolution, e.g. by using `plot(res, xlim=c(1000,1300))`.

3. For humans, how long (in years) would it take for a neutral mutation to fix?

□

Exercise 3.10. We have introduced the neutral fixation probability and fixation time for an ideal Wright-Fisher population. How do we need to adjust the results if we think of a natural population and the various concepts of effective population sizes?

4 Inbreeding and Structured populations

In the last section, we have seen that properties of natural populations can often be described by the theory of an ideal population if we use the concept of the effective population size. In this and the next chapter we will see that there are limits to this concept. In particular, in this chapter, we will focus on the effects of population structure.

A crucial property of the ideal Wright-Fisher model that we have used so far was the assumption of random mating, or panmixia. In terms of the coalescent, this assumption means that any two alleles (or haploid "individuals") from the offspring generation have the same probability to find a common ancestor in the previous generation. Most obviously, this is usually not the case in natural populations. For example, a Viennese mouse from the west side of the Danube will most likely mate with a mouse from the same side, and less likely with an east-side mouse. Similarly, lines of descent of mice from the same side of the river will coalesce earlier, on average, than lines of descent from mice from opposite banks.

A quick calculation shows that all these aspects are missed by the (inbreeding) effective population size $N_e^{(i)}$. Assume that there is a mouse population of (haploid) size $2N$, with N alleles on each side of the Danube. For simplicity, assume also that mice on each side mate randomly, while mice from opposite sides never mate. If we now pick two random individuals from the split population, there is a probability of $1/2$ that they are from the same side of the river, and in that case they are identical by descent with probability $1/N$. So overall, the probability for identity by descent is $1/(2N)$; we thus find $N_e^{(i)} = N$. This demonstrates that we need to introduce other concepts to capture the effects of population subdivision. In diploid populations, deviations from random mating are often measured as deviations from *Hardy-Weinberg equilibrium* which we will therefore explain first.

4.1 Hardy-Weinberg equilibrium

In the diploid Wright-Fisher model we assumed mating is random and a new diploid individual is formed by combining two random haploid gametes (taken from diploid parents). If two alleles A_1 and A_2 at a locus occur with frequencies p and $q = 1 - p$ we should find the following frequencies for the genotypes in the offspring generation:

$$\begin{array}{rll} p^2 & \text{for genotype} & A_1A_1, \\ 2pq & \text{for genotype} & A_1A_2, \\ q^2 & \text{for genotype} & A_2A_2. \end{array}$$

These are the so-called Hardy-Weinberg equilibrium frequencies of the genotypes. By measuring these frequencies in a natural population, we obtain a first test of whether the population fits to the null model of an ideal neutral population. Due to genetic drift, the match will never be perfect in a finite population (or a sample thereof), but a standard χ^2 test easily answers the question whether the differences are significant. Several other

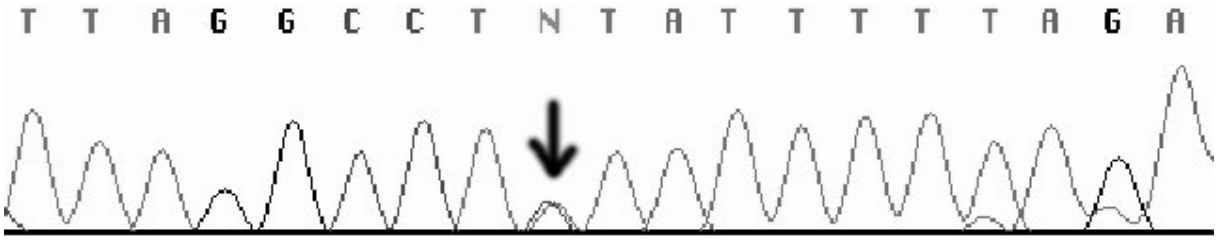


Figure 4.1: The output of a sequencing machine for the DNA of one individual. The machine recognizes the order of the bases and draws four curves for the amounts of *Adenine*, *Cytosine*, *Guanine* and *Tymine* it finds. At some places, as e.g. seen in the middle of the figure, there is a heterozygote meaning that at this site two bases are found by the machine. That means that there are two bases present in the individual, meaning that at this site it is heterozygous.

factors, including inbreeding, population subdivision, selection, and assortative mating may all cause a population to deviate significantly from Hardy-Weinberg proportions. However, an important point is that only factors that affect the population *in the present generation* will matter. In fact, whatever the distortion from Hardy-Weinberg equilibrium was in the parent generation, a single generation of random mating will restore the equilibrium in the offspring generation. As we will see later, this is in contrast to deviation from linkage equilibrium, which has a much longer memory (and therefore can tell us more about the history of a population).

Short detour: detecting heterozygotes

How do we distinguish homo- and heterozygotes after a sequencing reaction? To see this let us look at the output of a sequencing machine in Figure 4.1. The sequencer reads through the genome in four different channels, one for each of the four bases. These four channels are drawn as four different colors in the figure. At a homozygote site, only one channel is used and the base is easily identified by the color. But occasionally the sequencer finds two bases at a certain site (as at the position of the arrow in the figure). This is interpreted as a heterozygous site, where on one chromosome there is, e.g. a `T` (Tymin) whereas on the other there is a `C` (Cytosin).

Note that, although we can identify the heterozygous sites and also which bases form the heterozygote, it is impossible from the above graph to detect which chromosome carries which basepair. So when the individual is heterozygote for two sites, it is impossible to say which pairs of bases are on the same chromosome. This is usually referred to as not knowing *phase*. Not knowing phase is a big problem e.g. when one wants to estimate recombination rates. There are methods to find out phase, e.g. cloning but we will not treat this here.

4.2 Inbreeding

The mating of relatives to produce offspring is referred to as *inbreeding*. The most extreme form of inbreeding results from self-fertilization which is possible in many plants but also in snails and funghi. But also any other scenario that leads to higher relatedness of mated pairs than expected for random picks from the population will induce some level of inbreeding. The concepts of "population structure" and "inbreeding" are closely related. In fact, population structure with restricted migration between subpopulations will lead to inbreeding and we can think of the subpopulations as resembling large "families". Vice-versa, inbreeding is also possible without spatial population structure, simply as a consequence of the system of mating.

As an example, we consider the case of self-fertilization, where the analysis is relatively simple. Assume that in a diploid population fertilization can occur by either random mating or selfing, and that selfing occurs with probability p_s. We can now calculate first the probability that two homologous alleles *in a single offspring individual* derive from the same allele in the parent generation. This quantity is called the *inbreeding* coefficient f. We find

$$f = \frac{p_s}{2}$$

since we need, first, that both alleles are from the same diploid parent (which occurs with probability p_s), and, second, that both are copies from the same parental allele (probability 1/2). Let us calculate next the probability for identity by descent for two randomly picked homologous alleles in the offspring generation, i.e. the average single-generation coalescence probability. We find

$$p_{c,1} = \frac{1}{2N-1} f + \frac{2N-2}{2N-1}\frac{1}{2N} \approx \frac{1}{2N} f + \frac{1}{2N}., \tag{4.1}$$

which can be understood as follows: Two randomly picked alleles will be in one individual with probability $\frac{1}{2N-1}$ and in different individuals with probability $\frac{2N-2}{2N-1}$. In the first case, the probability to coalesce in one generation is f, in the other case it is $\frac{1}{2N}$. Rearranging the approximation in (4.1) gives the inbreeding effective population size under partial self-fertilization

$$N_e^{(i)} = \frac{N}{1+f} = \frac{N}{1+p_s/2}$$

Note that this result differs from the case of a subdivided population that we have discussed above. The reason is that inbreeding due to population structure (with fixed deme sizes) does not lead to a larger offspring variance. In contrast, it is easy to show that the offspring variance is enhanced in diploids with selfing (an allele in a selfing individual is likely the parent of two offspring alleles at once). Strictly speaking, it is thus offspring variance, not inbreeding that reduces the "inbreeding" effective population size[4]!

[4]At least with our definition, which follows Ewens (2004). Unfortunately, there are also various other definitions in the literature.

Exercise 4.1. To see the effect of inbreeding on genetic drift use the function `inbreeding()` of the `R`-package. This function simulates the time evolution of an allele A in a (partially) selfing population with inbreeding coefficient f. The population starts with two alleles (a and A) at equal frequencies. Drift makes one of the two go to fixation.

1. Consider several runs of the time evolution for some inbreeding coefficients. What do you assume to see for larger inbreeding coefficients? How can you observe a decreased effective population size?

□

Before we come to the next part of the exercise we introduce the mathematical notion of a *distribution function*.

Maths 4.1. *For any random variable X the distribution function of X is the function*

$$F_X(x) := \mathbf{P}[X \leq x].$$

This function increases and eventually reaches 1 Any point x_{med} with $F_X(x_{med}) = 0.5$ is called a median *of (the distribution of) X.*

2. Let's now look at the quantitative effect of inbreeding. Display for 200 runs the average fixation time. To do this use e.g. `plot(inbreeding(batch=200), what="fixed")`. This curve exactly tells you now in which proportion of the runs, one of the two alleles has fixed at a certain time. This means that you are actually displaying the distribution function - compare Maths 4.1 - of the fixation time under inbreeding.

   The median for the fixation time is the generation number for which 50% of all possible runs have already fixed. Compare the median for several coefficients. How well does your finding fit to the predicted change in the inbreeding effective size by a factor of $\frac{1}{1+f}$? To see this plot $(1+f) \cdot x_{med}$ for different values of f.

3. What is the largest effect inbreeding can have on the effective population size? Compare this to unequal sexratios.

Inbreeding and Heterozygosity

The inbreeding coefficient f was defined above as the probability for identity by decent. However, what we can measure from data is usually only the *identity by state*, i.e. whether two homologous alleles are identical. Identity by state is measured as *homozygosity* - or, equivalently, by its counterpart *heterozygosity*, which measures differences in allelic state. We have previously defined the heterozygosity H as the probability that two alleles from the population are different by state. In a structured and inbreeding population this definition can be refined:

- On the highest level, the *total heterozygosity* H_T is defined as the probability that two randomly chosen alleles from the entire population are different.

- On the lowest level, we define the *individual heterozygosity* H_I as the probability that two homologous alleles of a single, randomly chosen individual are different.

If there are two alleles A and a at the locus with frequencies p and $q = 1-p$, we obviously have $H_T = 2pq$. For a population in Hardy-Weinberg equilibrium, also $H_I = 2pq$, but in general H_I may deviate from this value. It turns out that this deviation is measured by the inbreeding coefficient.

Assume a population with inbreeding coefficient f. We pick one individual at random and denote the observed genotype by G. The two homologous alleles of this individual are either identical by descent (ibd) due to non-random mating or not ($\overline{\text{ibd}}$). In the former case, we can safely assume that they are also identical by state (where we exclude the possibility of a very recent mutation event). In the latter case, they are as closely related as two randomly chosen alleles from the population and are different by state with probability $2pq$. We calculate:

$$\begin{aligned}\mathbf{P}[G = AA] &= \mathbf{P}[G = AA|\text{ibd}] \cdot \mathbf{P}[\text{ibd}] + \mathbf{P}[G = AA|\overline{\text{ibd}}] \cdot \mathbf{P}[\overline{\text{ibd}}] \\ &= pf + p^2(1-f) = p^2 + fpq.\end{aligned}$$

Analogously we can calculate

$$\mathbf{P}[G = aa] = q^2 + fpq.$$

and so

$$\begin{aligned}\mathbf{P}[G = Aa] &= 1 - \mathbf{P}[G = AA] - \mathbf{P}[G = AA] \\ &= 1 - p^2 - q^2 - 2fpq = 2pq - 2fpq = 2pq(1-f).\end{aligned} \tag{4.2}$$

By rearranging (4.2), we obtain a new interpretation of the inbreeding coefficient f as the relative deviation of the total and individual heterozygosity,

$$f = \frac{2pq - \mathbf{P}[G = A_1A_2]}{2pq} = \frac{H_T - H_I}{H_T} \tag{4.3}$$

Thus f measures the proportion of genetic variation among homologous alleles in a population that is due to differences among individuals.

4.3 Structured Populations

An experimentalist who collects individuals in the field often takes samples from different places. She can suspect that the fact that she collected them from different places affects properties of the collected DNA patterns. So she must not forget about the locations of her samples. Let us look again at the sample from human populations which we already examined in example in Section 3.

Here a total of 41 sequences were taken from human X-chromosomes, 10 from Africa, 10 from Europe, 11 from Asia and 10 from America. To analyze structure in the model we have to define these groups. For the dataset this is already done. If the groups would not

have been defined yet, you can do it under `Data->Define Sequence Sets`. First, to get an overview over the amount of difference between the populations, we can do a pairwise analysis of divergence between populations. This shows how different they are.

Exercise 4.2. When comparing the African sample with the non-African one (this is also defined as a `sequence set`), `DNASP` outputs:

```
Between populations:
 Number of fixed differences: 0
 Mutations polymorphic in population 1, but monomorphic in population 2: 8
 Mutations polymorphic in population 2, but monomorphic in population 1: 6
 Shared Mutations: 2
```

1. From these numbers would you say that the two populations are very diverged? Or would you say that the whole sample is close to panmixia?

2. There is a certain amount of difference between Africa and the rest of the world. Between Africa and which subgroup (Europeans, Americans, Asians) is the most divergence as can be seen from these numbers?

□

We will use `Analysis->Gene Flow and Genetic Differentiation` to see more about the level of differentiation between the populations. One task when working with `DNASP` is to interpret the numbers on the screen. When executing `Gene Flow and Genetic Differentiation` we have to define the sequence sets that determine the subpopulations. Here we exclude `NonAfrican` because this is only the combination of Europeans, Americans and Asian sequences.

To make use of the geographic information of a sample, we need models that capture geographic space. These models are usually referred to as *structured models*. But also without these models we can make summary statistics that describe important facts about the population.

We saw that the inbreeding coefficient f can be seen as a measure that compares genetic variation within and between individuals. In the case of structured populations, Wright's fixation indices play an analogous role.

Fixation indices

If a population is structured into subpopulations, it is natural to define a third heterozygosity measure: H_S is the probability that two random alleles from the same, randomly chosen subpopulation are different by state. With three different levels of heterozygosity, H_T, H_S,

and H_I, we can define three so-called fixation indices or F-statistics (first introduced by Sewall Wright),

$$F_{IS} = \frac{H_S - H_I}{H_S}, \quad F_{IT} = \frac{H_T - H_I}{H_T}, \quad F_{ST} = \frac{H_T - H_S}{H_T}. \tag{4.4}$$

Note that all have the same form as (4.3). To derive the fixation indices from allele frequencies, we need mean values. Since we will need them often we should define them exactly:

Maths 4.2. *If $x = (x_1, \ldots, x_n)$ is a list of numbers the* mean *of the list is given by*

$$\overline{x} = \tfrac{1}{n} \sum_{i=1}^{n} x_i.$$

With this procedure, if e.g. another list $y = (y_1, \ldots, y_n)$ is given distinguish between

$$\overline{x}\,\overline{y} = \Big(\frac{1}{n} \sum_{i=1}^{n} x_i\Big)\Big(\frac{1}{n} \sum_{i=1}^{n} y_i\Big), \tag{4.5}$$

indicating the product of means and

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{4.6}$$

which is the mean of the product.

Assume data were collected from d locations and allele frequencies are $p_1, q_1, \ldots, p_d, q_d$ for two alleles A and a and $p_{Aa,j}$ is the observed frequency of heterozygotes in deme j. Then we define

$$\overline{p} = \tfrac{1}{d} \sum_{j=1}^{d} p_j, \qquad \overline{pq} = \tfrac{1}{d} \sum_{j=1}^{d} p_j q_j, \qquad \overline{p_{Aa}} = \tfrac{1}{d} \sum_{j=1}^{d} p_{Aa,j},$$
$$H_S = 2\overline{pq}, \quad H_T = 2\overline{p}\,\overline{q}, \quad H_I = \overline{p_{Aa}}.$$

The three measures F_{IS}, F_{IT}, and F_{ST} detect differences in genetic variation due to non-random mating at different levels. The most widely used index in the case of population structure is F_{ST}. It serves as a measure for the proportion of genetic variation among individuals drawn from all subpopulations that is due to genetic differences between subpopulations. As an exercise, we calculate F_{ST} from the human X chromosome dataset.

Exercise 4.3. In `Gene Flow and Genetic Differentiation` we saw for our data that several numbers were computed. Among them, also F_{ST}. In the help of `DNASP` it is mentioned that F_{ST} is calculated according to HUDSON *et al.* (1992). In Figure 4.2 you see what is actually calculated there. Assume you are given sequence data of length one nucleotide, from two demes, 5 individuals per deme. The data is `A, A, T, T, T` for deme 1 and `A, A, A, A, T` for deme 2.

Data analysis: For each sample, F_{ST} was estimated by

$$\langle F_{ST} \rangle = 1 - \frac{H_w}{H_b} \qquad (3)$$

where H_w is mean number of differences between different sequences sampled from the same subpopulation, and H_b is the mean number of differences between sequences sampled from the two different subpopulations

Figure 4.2: The definition of F_{ST} taken from (HUDSON *et al.*, 1992).

- What is the mean number of pairwise differences within subpopulation 1 and withing subpopulation 2?
- What is the mean number of pairwise differences for all sequences?
- What is $1 - \frac{H_w}{H_b}$?
- Compute F_{ST} as given above.
- Do the definitions of F_{ST} and the one given by Hudson match? If not, where is the difference?

□

An important question for anyone using F-statistics is: in which range are the values of the fixation indices? The answer to this question gives a hint which F-values are expected in structured populations. By definition all of the F-coefficients are smaller than 1. The value of F depends on whether there an excess or a deficiency in heterozygotes in the subpopulations compared to the total population. If the subpopulations are in Hardy-Weinberg equilibrium, there is always an excess of homozygotes in the subpopulations. This is not only a theoretical prediction, but can be computed as we will see next.

We can write

$$\overline{pq} = \frac{1}{d}\sum_{i=1}^{d} p_i(1-p_i) = \frac{1}{d}\sum_{i=1}^{d} p_i - \frac{1}{d}\sum_{i=1}^{d} p_i^2 = \overline{p} - \overline{p^2}$$

Observe here the difference between $\overline{p}^2$ and $\overline{p^2}$, i.e.

$$\overline{p}^2 = \Big(\frac{1}{d}\sum_{i=1}^{d} p_i\Big)^2, \qquad \overline{p^2} = \frac{1}{d}\sum_{i=1}^{d} p_i^2.$$

Actually F_{ST} can be seen to be related to the variation in the p_i's. This is measured by the sample variance.

Maths 4.3. *If $x = (x_1, \ldots, x_n)$ is a list of numbers the* sample variance *of the list is given by*

$$\tilde{s^2}(x) = \tfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \tfrac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \tfrac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\overline{x}\tfrac{1}{n}\sum_{i=1}^{n} x_i + \overline{x}^2 = \overline{x_i^2} - \overline{x}_i^2.$$

Note that by the first equality $\tilde{s^2}(x) \geq 0$ and it equals 0 if and only if all values in the list are the same.

Therefore

$$H_T - H_S = 2(\overline{p}\,\overline{q} - \overline{pq}) = 2(\overline{p} - \overline{p}^2 - \overline{p} + \overline{p^2}) = 2(\overline{p^2} - \overline{p}^2) = 2\tilde{s^2}(p) > 0$$

and thus

$$F_{ST} = \frac{\tilde{s^2}(p)}{\overline{p}\,\overline{q}} \quad \text{and so } F_{ST} > 0$$

where $\tilde{s^2}(p)$ is the sample variance of the p_i's. So from these calculations there cannot be a deficiency of homozygotes in the subpopulations compared with the total population.

There is a connection between the three fixation indices F_{IS}, F_{ST} and F_{IT}. So two of them always determine the third one. The connection is

$$(1 - F_{IS})(1 - F_{ST}) = \frac{H_S}{H_I}\frac{H_T}{H_S} = \frac{H_T}{H_I} = 1 - F_{IT}. \tag{4.7}$$

To see in which range F_{IS} can be found consider the most extreme case that there are only heterozygotes in all subpopulations. Then $p_i = \frac{1}{2}$ for all demes i and thus

$$F_{IS} = \frac{2\overline{pq} - \overline{p}_{12}}{\overline{pq}} = \frac{\frac{1}{2} - 1}{\frac{1}{2}} = -1$$

which is the lower bound for F_{IS}. With (4.7) it is clear that also F_{IT} is between -1 and $+1$.

As F_{ST} measures the amount of the excess of homozygotes a small F_{ST}-value means that our data comes from a population that is unstructured or only a little bit structured. What a *small* value for F_{ST} is depends also on the species. For *Drosophila* an F_{ST} of 0.2 would be considered big whereas for plants it would not.

The Wahlund effect

As we have seen in the last paragraph, even when all subpopulations are in Hardy-Weinberg equilibrium, there will be an excess of homozygotes in the total population - unless allele frequencies are exactly the same in all subpopulations. This is called the *Wahlund effect* after its inventor Sten Wahlund. The consequece is that data from a subdivided population look like data from an inbreeding population.

Exercise 4.4. Suppose you have sampled individuals from a plant population and you look at one gene that has two alleles A_1 and A_2. You find 28 individuals that have genotype A_1A_1, 32 that have A_1A_2 and 40 that have A_2A_2.

1. What is the frequency of A_1 in the whole population?
2. What would the Hardy-Weinberg equilibrium be?
3. Does this plant population deviate from Hardy-Weinberg?
4. What would be your conclusion about this population?

Now imagine there are two subpopulations (or demes) of this plant. They grow not far apart, but for some reason pollinators just don't fly from one population to the other. Now suppose you would take samples from both populations and would find the following number of individuals:

	A_1A_1	A_1A_2	A_2A_2
Deme 1	26	13	1
Deme 2	2	19	39

1. Answer the above questions in this case.
2. What would the value of F_{ST} be in the above example?

□

F_{ST} already gives some idea about the amount of structure. There are also statistical tests to decide whether the population shows structure. Look at the table of Exercise 4.4. This is usually referred to as a *contingency table*. In this instance it has two rows and three columns. Extending it we could also write

	A_1A_1	A_1A_2	A_2A_2	$\sum$
Deme 1	26	13	1	40
Deme 2	2	19	39	60
$\sum$	28	32	40	100

where the row and colum sums are added. Assuming that the genotypes are equally distributed among the demes we would e.g. expect that in deme 1 genotype A_1A_1 has a total number of $100 \cdot \frac{40}{100} \cdot \frac{28}{100} = 11.2$. However we observe 26 in this group. A χ^2-test can be used to decide if the deviance of the table from a random distribution of the individuals to the groups is significant. The test statistic has the form

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

where the sum is over all cells in the table. This test statistic is approximately χ^2 distributed with $(R-1)(C-1)$ degrees of freedom, R denoting the number of rows and C the number of columns of the contingency table.

Exercise 4.5. In the analysis `DNASP` performs a χ^2 test to infer structure. However it uses not genotypes but haplotypes. Do the `Gene Flow and Genetic Differentiation` using the samples of Africa and the rest of the world.

1. How many haplotypes do you find in the total population, how many in Africa and how many outside? How many are shared between Africa and Non-Africa?

2. For the χ^2 test `DNASP` outputs

```
Chi-square (table), Chi2: 31,239   P-value of Chi2: 0,0031 **;   (df = 13)
   ns, not significant;  *, 0.01<P<0.05;  **, 0.001<P<0.01;  ***, P<0.001
```

   Why is `df` (which is the number of degrees of freedom) 13? With this result, can you conclude if there is genetic differentiation between Africa and the rest of the world?

□

4.4 Models for gene flow

Although we can now detect a deviance from panmixia by the use of the F-statistics and a χ^2-test we have no model yet to explain these deviances. The oldest and most widely used models of population structure are the *mainland-island* and the *island* model. In these models, it is assumed that the subpopulations have existed for a long time and migration rates are constant over time, so that an equilibrium between migration and drift is established. If these assumptions are met, we can use the F-statistics to estimate migration rate.

Several schemes of gene flow are possible mainly depending on the sizes of the subpopulations and the connections between the subpopulations; but only in very special models we can calculate quantities of interest, such as F_{ST}. The two examples we cover are the *mainland-island* and the *island model*. The first is a model for a population consisting of one big subpopulation and one (or more) smaller ones; in the second model the population consists of many small subpopulations, on islands, that exchange genes. These two models

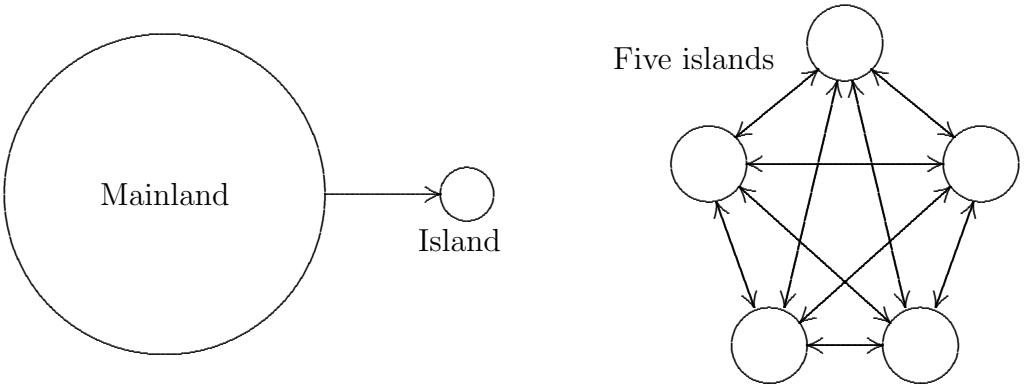


Figure 4.3: The mainland-island and the island model for gene flow. In the mainland island model only migration from the mainland to the island plays a role, whereas for the island model gene flow occurs between any two demes.

can be found schematically in Figure 4.3. In each case, the subpopulations are assumed to be randomly mating. You may be surprised that these models deal with islands, but the idea of subpopulations on islands is not as strange as you might think. Island in this case can be seen as islands of suitable habitat in a sea of unsuitable habitat. Think of deer in Europe that live in islands of forest surrounded by seas of agricultural land and cities.

The mainland-island model

Consider a population living on a continent and occasionally sending migrants to surrounding islands. For simplicity assume that there is only one island; an allele A has frequency p_m on the continent (where the m stands for *mainland*) and p on the island. The continent is so big that drift is negligible and migrants that come from the islands back to the continent can be ignored. A fixed fraction m of the island population is replaced every generation by individuals from the mainland. Ignoring random sampling, i.e. ignoring genetic drift the allele frequency follows the deterministic equation

$$p_{t+1} = (1-m)p_t + mp_m = p_t + m(p_m - p_t)$$

which has

$$p = p_m$$

as a stable equilibrium.

The island model

Quite different is the island model. Here a population is distributed on several islands. On each island a fixed fraction of the population is replaced by individuals from the so-called gene-pool. The genepool consists of all individuals on all islands. Here the allelic frequencies can again be computed. As an allele taken from one subpopulation has the

chance $1 - m$ to be a descendant from the same subpopulation and m to come from somewhere else we have, again ignoring genetic drift,

$$p_{t+1} = (1 - m)p_t + m\overline{p} = p_t + m(\overline{p} - p_t).$$

As we assume no drift $\overline{p}$ stays constant and is therefore also the equilibrium frequency of A.

Theoretical predictions on the excess of homozygotes

But what can we say about the excess of homozygotes in these two models? Let us assume that we are dealing not only with 2 but with an arbitrary number of alleles and every mutation will create a new allele. These are the assumptions of the *infinite alleles model*. Furthermore, for the mainland-island model assume that the mainland, which is much bigger than the island, is not affected by drift. Backwards in time that means that just picking two chromosomes at random from the same population they will not coalesce for a very long time. During this very long time it is very likely that a mutation has happened and so any two randomly picked chromosomes from the mainland will be different. As the mainland is much bigger than the island, this means that any two picked chromosomes from the population are different.

For the island model, assume that there are a lot of islands, so that picking two individuals at random means that they will come from different islands. Furthermore, because the number of islands is very high the population as a whole is also not affected by drift. Hence similar to the mainland-island model, two chromosomes from different demes are very likely to be different.

This means that under these assumptions we have

$$H_T \approx 1.$$

And what about H_S? Well, let us consider here two chromosomes from one deme and think about their most recent common ancestor. Looking backward in time two events are possible, migration away from the deme or coalescence of the two lines. Migration away from the deme occurs when forward in time one lineage was in fact a migrant from somewhere else. As m is the migration probability the first time T_m until one of the two lineages migrates away from the deme is geometrically distributed with parameter $2m$. Additionally the time when the two times coalesce T_c is geometrically distributed with parameter $\frac{1}{2N}$. As the geometric distribution is close to the exponential distribution with the success probabilities as rates (compare Maths 2.4) we can compute the probability for a heterozygote as the probability that one lineage migrates before both coalesce. To do this we use Maths 3.1.

This means that approximately

$$H_S \approx \mathbf{P}[T_m < T_c] = \frac{2m}{2m + 1/(2N)} = \frac{4Nm}{4Nm + 1}.$$

So we can conclude that

$$F_{ST} = \frac{H_T - H_S}{H_T} \approx 1 - H_S = \frac{1}{4Nm + 1}. \tag{4.8}$$

Let us define

$$M = Nm,$$

which is the total number of (diploid) migrants per deme in each generation. The haploid number is then $2M$. As we have seen we can compute F_{ST} from data. Using (4.8) we can therefore - under the island model of migration - estimate M from data.

Exercise 4.6. Can you give an estimate of M from (4.8)? Calculate this estimator for the human X chromosome data set. Discuss whether the assumptions of this estimator are met for the human population. □

Exercise 4.7. Use the function `popSubdivision()` from the `R`-package with

```
>ret<-popSubdivision(N=50,stoptime=500,m=0,demes=10,init.A=0.5,mainland=FALSE)
```

Plot the frequencies in all demes using `plot(ret)`.

You are now looking at an island model population with 10 demes. The demes are really populations in themselves because migration is 0, but for now we will keep calling them demes. Setting the migration rate to 0 means that the demes are completely independent of each other. The initial frequency of the A allele is 0.5 . And we run the model for 500 generations. (If the mainland parameter is `FALSE` we have an island model if it is `TRUE` we have a mainland-island model.) Because the individual populations are small drift has a strong effect on them. The plot of allele frequencies shows you the frequency of the A allele in each of the demes. You should see 10 lines on your plot.

1. After 100 generations how many of the populations are fixed for either of the two alleles? Run the model a couple of times to see if the answer is always the same.

2. Now you can allow for migration. You can do this by choosing `m = 0.01`, which gives every individual the probability of 0.01 to be replaced by a migrant. The probability that a migrant has the A allele is simply the overall frequency of A. What is the mean number of migrants per island per generation?

3. Now, after 100 generations how many of the populations are fixed for either of the two alleles? Run the model a couple of times to see if the answer is always the same. Can you explain your observation? What does migration to variation?

4. Now try increasing m in small steps. What happens to your demes? How high do you need to make m to lose all (detectable) population structure? Note that setting m to 1 makes all individuals migrants and this is the same as saying that all demes are really one population.

□

Non-equilibrium approaches to population structure

In the last years, programs such as `BAPS` and `Structure` became important in the analysis of population structure. These programs are not based on F-statistics and they are not based on migration drift equilibrium assumptions. They do assume Hardy-Weinberg equilibrium in the subpopulations. Another program, `BayesAss` does not make the Hardy-Weinberg assumption. We will discuss the approaches briefly in the lecture.

5 Genealogical trees and demographic models

Look again at data from the human X chromosome. If you assume that they are consistent with the neutral theory you can now, for example, estimate the population size. But how can you know that the data is consistent with the neutral theory? You might not be able to answer this question at the moment, and well for the following reason: we haven't yet talked about what patterns to expect in sequence diversity, so we have no predictions and so nothing we can test. For example, we have looked at θ and how we can use it to estimate the mutation rate or the population size. But the neutral theory doesn't tell us what value θ should have, and so a single θ value doesn't tell us whether our data is consistent with an ideal neutrally evolving population. Now we will look at two aspects of a set of DNA sequences that are more complex and that can be used to test whether the data are consistent with neutrality. The two aspects that we will look at are the *site frequency spectrum* (in Subsection 5.2) and the *mismatch distribution* (in Subsection 5.4). The reason that we look at exactly these two aspects of the data is that they are independent of the mutation rate and population size. We will also, in Subsection 5.3, consider the effect of changing population size on the site frequency spectrum. However, before we go into all this we will need to have a closer look at genealogical trees. Two things to keep in mind at all times are that (i) processes in the population are stochastic (for example where and when mutations happen and which individual has how many offspring) and (ii) we are usually looking at only a small random sample taken from the whole population.

5.1 Genealogical trees

The coalescent, which was introduced in Section 2, is the ideal candidate for an approach to derive predictions under the neutral theory. That is because it deals with *samples* from populations and not with complete populations, which reduces the complexity of our studies. Furthermore, the coalescent can deal with the most important aspects of the history of a sample: the mutations that happened and the genealogical relationships of the sampled sequences. These genealogical relationships are usually represented in a tree. Times when mutations happen are determined by the mutation rate and times when coalescences happen are determined by the (effective) population size. As you will see later, the coalescent can also deal with recombination and migration and with changing population sizes.

The history or genealogy of a sample has many different aspects. The most obvious are genealogical relationships, which are usually represented in a tree. The second is the lengths of the branches between the nodes of the tree. These lengths can be given just in number of mutations (because that is the information we usually have) but the lengths can also be given in generations (which allows us to predict the number of mutations on the branch). Then, another aspect of a tree - which we don't observe but infer -is the intermediate sequences, including the sequence of the MRCA, and the exact timing of the mutations.

Exercise 5.1. The following sequences are from 5 individuals, the alignment is already done.

```
1  AATCCTTTGGAATTCCCT
2  GACCCTTTAGAATCCCAT
3  GACCCTTTAGGATTCCAT
4  GACCTTCGAGAGTCCTAT
5  GACCTCCGAGAATCCTAT
```

Assume that in the history of your sample every mutation hits a different site. Can you draw a bifurcating tree, with mutations on the branches, that would produce the observed data?

Reconstructing trees, or inferring phylogenies, is a big scientific field of its own, we will not spend too much time on it in this course. Look at FELSENSTEIN (2004) if youo want to learn more about this topic. At least, from the last exercise you can see that the topology of the genealogical tree can, at least in certain cases, be reconstructed from data.

As you might have noticed, mutations (SNPs in this case) can be divided in two classes, those relevant and those irrelevant for the tree topology. Mutations that split the sample in 1 and 4 (or 1 and $n-1$ if the sample has size n) tell us nothing about the topology whereas every other mutation divides the sample in two subsamples that are bigger than 1 and that are separated by the mutation. In DNASP you can view these sites by clicking on `Parsimony Informative Sites` on the `view data` sheet.

Exercise 5.2. In the tree that you have made in the last exercise where are the parsimony informative sites and where are the parsimony uninformative sites?

The MRCA sequence and the outgroup

Every sample of homologous sequences must have a common ancestor. Now that we know the genealogical tree, we can ask whether we can also infer the sequence of the MRCA of the sample? If you try this for all sites in the above example, you will see that you can infer the sequence for some but not all sites. E.g. the first mutation divides the sample in those carrying an `A` and those carrying `G`, but we don't know whether the MRCA carried an `A` or a `G`

There is a way to find out the whole sequence of the MRCA, with relative certainty by using an outgroup sequence. This means that you add to your tree a lineage where you can be sure - for which reason ever - that the MRCA of the extra lineage with the whole sample lies further in the past than the MRCA of the sample. This lineage can either be a sequence of an individual sampled from a different population or from a different species. The species you use for the outgroup should not be too far away (i.e., the MRCA of the outgroup and the rest of the sample should not be too far in the past) because then the assumption that every new mutation hits a different site in the sequence is likely

to be violated. The outgroup for human data is usually a chimpansee sequence, and for *Drosophila* data it would be another *Drosophila* species.

Exercise 5.3. Assume that the homologous sequence of a closely related species for the data of Exercise 5.1 is

Outgroup `AACCCTTTAGAATTCCAT`

Draw the tree topology for the sample including the outgroup. Can you now say what the sequence of the MRCA of the sample is?

By using the outgroup sequence we found the point in the tree which is furthest in the past. This specific point is often called the *root* of the tree.

Exercise 5.4. Genealogical trees are important to compare observations in data with theoretical predictions.

1. Look at the human X chromosome data using `DNASP`. Forget all models you learned in the past sections. Can you find some numbers, also called statistics, (such as e.g. the number of segregating sites) that describe your data?

Now think again of genealogical trees. Take e.g. the number of segregating sites and assume you know μ and N. Then we can drop mutations that lead to segregating sites under the infinite sites assumption on the coalescent. This means that the probability of a given number of segregating sites can be computed by the coalescent process.

2. Use the data from Exercises 5.1 and 5.3. Here you already know something about the genealogical trees behind the data. Take the statistics you found in 1. Can you compute them not by looking at the data but by looking at the genealogical tree and the mutations that hit the tree?

3. Take your statistics you found in 1. Here comes a conjecture:

   > The distribution of every statistic that can be computed from polymorphism data alone can at least principally be calculated from the distribution of genealogical trees using a mutation process.

   In other words: the probability that the statistic takes a certain value is a function of the mechanism of creating the genealogical tree including the mutations.

   Can you agree with this conjecture? Or can you falsify it using your statistics from 1?

Number of possible rooted and unrooted trees

It is hard to be sure that you have found the best tree for the above dataset, even if you are pretty sure that there is no better tree than the one you have found. Now, we have not really defined what a good tree is and there are different ways to do that. But more problematic is usually the number of possible trees. Even if we only consider the topology of a tree (not the branch lengths or the mutations etc.), *tree space* is multidimensional and very large. Let us calculate the number of possible trees for a given sample size:

Given a tree with n leaves (which corresponds to n sampled sequences) there must be $n-1$ coalescence events until the MRCA of the sample. This creates $n-1$ vertices in the tree. This makes a total of $2n-1$ vertices (leaves or internal vertices) in the coalescent tree. But how many branches are in this tree. Every vertex has a branch directly leading to the next coalescence event. Only the MRCA, which is also a vertex in the tree does not have a branch. This makes $2n-2$ branches in a rooted tree with n leaves. As two branches lead to the root, i.e. the MRCA the number of branches in an unrooted tree with n leaves is $2n-3$.

Let B_n be the number of topologies for unrooted trees with n leaves. Assume you have a tree with $n-1$ leaves, which represent the first $n-1$ sampled sequences. In how many ways can the nth sequence be added to this tree. Any branch in the tree can have the split leading to the nth leave. As there are $2n-3$ branches in a tree with n leaves there must be $2n-5$ branches in a tree with $n-1$ leaves. This gives

$$\mathcal{B}_n = (2n-5)\mathcal{B}_{n-1}.$$

This is a recursion as the nth number is given with respect to the $n-1$st. However in this case it is easy also to give an explicit formula for $\mathcal{B}_n$ as

$$\mathcal{B}_n = (2n-5)\mathcal{B}_{n-1} = (2n-5)(2n-7)\mathcal{B}_{n-2} = \dots.$$

This must be

$$\mathcal{B}_n = 1 \cdot 3 \cdots (2n-7) \cdot (2n-5).$$

Of these trees only one represents the true history of your sample.

Exercise 5.5. Let us see what this number of trees really is and how big it can be.

1. How many tree topologies for unrooted trees are there for a tree with 4 leaves? Can you draw all of them?

2. A usual sample size is 12. How many tree topologies do you obtain in this case? A large sample would include 20 sequences. How many tree topologies do you count then?

   HINT: You can use `R` to compute this number.

Simple estimation of branch lengths

Given that we have a tree topology of our sample, with all the mutations and the sequence of the MRCA, the next thing we can look at s the branch lengths. Only then we can claim to have reconstructed the genealogical tree.

We can use the number of mutations between two nodes to estimate the branch lengths. Obviously when we find many mutations on a branch it would be natural to assume that this branch is a long branch. Let us first do some reverse engineering and assume we know branch lengths in numbers of generations in an ideal Wright-Fisher model. Assume a branch has length L. As in each generation the per locus mutation rate is μ, the probability that we find k mutations on this branch is

$$\mathbf{P}[k \text{ mutations on branch of length } L] = \binom{L}{k}\mu^k(1-\mu)^{L-k}.$$

As typically L is large compared with k and μ is small we can approximate this probability by using the Poisson-distribution as we already did in Maths 2.1. The parameter of the Poisson distribution is μL and so

$$\mathbf{P}[k \text{ mutations on branch of length } L] \approx e^{-\mu L}\frac{(\mu L)^k}{k!}.$$

As the expectation of a Poisson-distribution equals its parameter we have

$$\mathbf{E}[\text{number of mutations on branch of length } L] = \mu L = \frac{\theta L}{4N}.$$

So the easiest estimator of the branch length is $L = \frac{S}{\mu}$ where S is the number of SNPs. Recall that branches are supposed to be short near the tips of the tree and long towards the root. In this simple estimate of the branch length we have completely ignored this knowledge. But with a little more sophisticated calculations we could make a much better estimate of the branch lengths.

5.2 The frequency spectrum

When you reconstructed the tree in Exercise 5.1 you already observed that mutations can be ordered by the split they produce in the sample. E.g. the first mutation in the exercise split the sample in 1/4. By using the outgroup from exercise 5.3 we know that state `A` is ancestral (because it is also carried by the outgroup) and `G` is derived. We say that this mutation has size 4, because there are 4 sequences with the derived state.

Exercise 5.6. Use the sequences from exercise 5.1 and the genotype of the MRCA which you found out in Exercise 5.3. How many mutations are there of size 1, 2, 3 and 4? How many mutations of size 5 do you expect?

Draw a histogram of the sizes of mutations, i.e. on the X-axis you put the size of the mutation and the Y-axis is the count of the number of mutations with this size in the data.

The last figure you produced is called the *site frequency spectrum* and is often used to infer historical events in the population from data. Our next task is to find a theoretical prediction what the frequency spectrum looks like under a neutral model of a constant size population. Therefore denote by S_i the number of mutations in your sample that are of size i.

For this we have to have a closer look at the coalescent trees. We say the tree is in state k when it has currently k lines (so it always starts in state n and ends in state 2). A branch at state k is of size i if exactly i of the sequences are descendants of this branch. So in order to get the expected number of mutations of this size we have to sum over all possible states (2 to n) and within each state we sum over all branches (1 to k) the probability that a branch has size i times the expected number of mutations on that branch. Then,

$$\mathbf{E}[S_i] = \sum_{k=2}^{n} \sum_{l=1}^{k} \mathbf{P}[l\text{th branch at state } k \text{ is of size } i] \cdot \mathbf{E}[\text{number of mutations on } l\text{th branch at state } k].$$

The second term in this sum is easy because we are only calculating expectations. We have here

$$\mathbf{E}[\text{number of mutations on } l\text{th branch at state } k] = \mu \cdot \mathbf{E}[\text{length of the } l\text{th branch at state } k] = \frac{2\mu N}{\binom{k}{2}} = \frac{\theta}{k(k-1)}$$

as the state k has a length $\frac{2N}{\binom{k}{2}}$ in expectation.

The first term is more tricky. There is a way to generate the tree topology from the root of the tree to its leaves which turns out to be a useful idea. This is done using the *Polyas urn scheme.*

Maths 5.1. *A* Polya urn scheme *is the following: Take an urn conatining balls, namely k balls with k different colors. Take out one ball, put it back to the urn and add one ball of the same color. Do this again. And again...*

Assume a tree of state k. The k balls in the urn represent the k branches at that state. Generating the $k+1$ state of the tree from the kth means picking a branch at random and splitting it. In the urn this amounts to adding a ball of the same color. Same colors here mean that the ancestor at the 'beginning of the urn scheme', i.e. at state k is the same. From the $k+1$st to the $k+2$nd the same game begins. Every branch is equally likely to split and this is exactly also done by Polyas urn.

So when starting the Polya urn with k balls and stopping it when it has n balls the number of balls of each color represents the number of individuals in the sample who are descendants from a specific line.

Now consider the ℓth branch at state k. For this consider the Polya urn that starts with k colors and consider the ℓth one. We must put $n-k$ balls into the urn to obtain the

sample of size n. Exactly when we put $i-1$ balls of color ℓ into the urn the ℓth branch will have i offspring among the sample of size n. Let us consider an example with $k = 5, n = 10$ and $i = 4$. We write e.g.

$$(\ell \,\square\square\, \ell \,\ell)$$

Here ℓ stands for a ball of color ℓ and $\square$ stands for a ball of any other color. In this example we start putting color ℓ, then a different color (i.e. one of the $k-1$ other colors), again a different one, and then two balls of color ℓ. As three balls of color ℓ enter the urn this leads to $i = 4$. The probability for a configuration like this is

$$\begin{aligned}\mathbf{P}[\ell \,\square\square\, \ell \,\ell] &= \frac{1}{k}\cdot\frac{k-1}{k+1}\cdot\frac{k}{k+2}\cdot\frac{2}{k+3}\cdot\frac{3}{k+4}\\ &= \frac{(i-1)!(k-1)\cdots(n-i-1)}{k\cdots(n-1)}\end{aligned}$$

as exactly $i-1$ balls of color ℓ and $n-k$ balls of different colors enter the urn. (Check this for the above example.) This is true for any configuration that contains $i-1$ balls of color ℓ and $n-k-i+1$ others. There are $\binom{n-k}{i-1}$ of these configurations as we only have to distribute the $i-1$ balls of color ℓ to the $n-k$ slots of the configuration. Altogether this gives

$$\begin{aligned}\mathbf{P}[l\text{th line at state } k \text{ is of size } i] &= \binom{n-k}{i-1}\frac{(i-1)!(k-1)\cdots(n-i-1)}{k\cdots(n-1)}\\ &= \frac{k-1}{i}\binom{n-k}{i-1}\frac{i!}{(n-i)\cdots(n-1)} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{k-1}{i}.\end{aligned}$$

Exercise 5.7. A coalescent with 5 individuals can have several topologies. Look at the split the root of the tree generates in the sample. What is the probability that 1,2,3 or 4 leaves lie on one side of the root with the rest on the other side?

A little more maths with binomial coefficients:

Maths 5.2. *There are many formulas for these binomial coefficients. The most simple one is*

$$\begin{aligned}\binom{n}{k-1}+\binom{n}{k} &= \frac{n!}{(k-1)!(n-k+1)!}+\frac{n!}{k!(n-k)!} = \frac{n!k+n!(n-k+1)}{k!(n-k+1)!}\\ &= \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}\end{aligned}$$

This gives

$$\begin{aligned}\mathbf{E}[S_i] &= \sum_{k=2}^{n}\sum_{l=1}^{k}\frac{\binom{n-k}{i-1}}{\binom{n-1}{i}}\frac{k-1}{i}\frac{\theta}{k(k-1)}\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\sum_{k=2}^{n}\binom{n-k}{i}\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\sum_{k=2}^{n}\Big(\binom{n-(k-1)}{i}-\binom{n-k}{i}\Big)\\ &= \frac{\theta}{i}\frac{1}{\binom{n-1}{i}}\Big(\sum_{k=1}^{n-1}\binom{n-k}{i}-\sum_{k=2}^{n}\binom{n-k}{i}\Big) = \frac{\theta}{i}.\end{aligned}$$

So under the neutral model with constant population size the prediction is that among all mutations those of size i have a relative frequency of $\frac{1}{i}$.

Exercise 5.8. Let us check how good our theoretical prediction for the site frequency spectrum is reflected in simulations. To do this we use `seqEvoNeutral()` from the `R`-package.

1. Use the option `wait=-1, sfs=TRUE` to see the joint evolution of the sequences and the site frequency spectrum. From the starting configuation, you can obtain the site frequency spectrum by hand using only the sequences on the left side.

2. Given `stoptime=100`, in how many generations do you see singletons? In how many genertions are there variants with frequency $N-1$? Can you explain your findings?

Exercise 5.9. What do you think, using the frequency spectrum, do the sequences from exercise 5.1 coincide with the neutral model of constant size?

Exercise 5.10. `DNASP` can also display the frequency spectrum. As we saw there are mutations the split the sample in k and $n-k$ while others split in $n-k$ and k derived and ancestral states. Without knowing the sequence of the MRCA we have no means to distinguish these two classes. Thus, `DNASP` puts these two classes together.

There should be enough polymorphic sites to see something in the frequency spectrum. In Gavrilin *et al.* (2000) two sets of Polioviruses were studied. The data is stored in the file `PolioVP1.nex`. Let us exclude some lines from the sample. One line from Azerbeijan `169AZB59` was already sampled in 1959 and the last 5 lines which were sampled in the USA. (Use `Data->Include/Exclude Sequences`.) The rest forms one set of Polioviruses which are highly polymorphic. Using `Analysis->Population Size Changes` you can plot the frequency spectrum.

1. The X-axis only reaches until 11 although 23 sequences were used. Why is this the case?

2. Do you agree with the opinion of the authors that in this group mutations are predominantly neutral?

3. Which assumptions on the model are possibly not met for the data? Could they confine your finding in the last question?

The frequency spectrum gives a nice graphical summary of the data that can easily be compared with a theoretical prediction. However this expectation should not be taken too literally. Data collected in population genetics are almost always correlated in a statistical sense. So sampling more sequences does not mean that the observed frequency spectrum should be closer to the expectation.

Let us make an example to illustrate this: Assume an unfair dice. Unfair means that the 6 numbers do not occur with equal probabilities. However you do not know how unfair it is. You can only assign probabilities on how big the deviance from fairness is. It might well be that still on average you through a 3.5 as for a fair dice. An extreme case would be that with probability $\frac{1}{6}$ the dice the shows 1 with 100% certainty, and also with $\frac{1}{6}$ it shows 2 with certainty and so on. Throwing this dice imposes correlations on the random outcomes of throwing the dice. Take the above example: once you know the first result of rolling the dice you can already predict all others because you know that with certainty only one number is adopted. After a long run of trials with the dice you will not have reached to 3.5 on average which is because these rolls were correlated.

But what does this mean for sequence data? Well, suppose that you have found unexpectedly many singletons in your sample. And you now suddenly have extra sequences from the neighboring locus. Analyzing the new data you find again many singletons. At first glance you may think that this is additional proof that something strange is going on in the population. However this need not be the case. The results from the two neighbouring loci are very much correlated. The second locus gives you no independent evidence of anything, you have only added more of the same information. In order to know more of the population you need information from unrelated loci, for example from another chromosome.

Nevertheless, the frequency spectrum is a useful tool. It is especially useful because its predictions do not depend on the mutation rate or the population size. The frequency spectrum - as already indicated by DNASP - is often used to infer changes in population size. In order to understand how it can be used for that we nee to look at some models for changing population sizes.

5.3 Demographic models

One main property of the standard Wright-Fisher model is a constant population size. Here, we will relax this assumption, i.e., we consider models with varying population sizes. Several demographic events lead to patterns in SNP data that can be detected. As you learned in Exercise 5.4 this is mainly because they highly affect what genealogical trees look like.

But before we will speak about the coalescent backward in time let us talk about the forward in time evolution. Assume the population size at time t is N_t and assume we know these numbers for any point in time. Time is measured in generations. To adapt the Wright-Fisher model to the case of a fluctuating size we condition it on the path of population sizes N_t. So how is the generation at time $t+1$ built from the tth generation? One feature of the Wright-Fisher for constant population size was that we can model the ancestry of the next generation by simply saying that every individual in the generation $t+1$ chooses its ancestor at time t purely random; furthermore all individuals in the population choose their parents independently. This can still be done for a non-constant population size.

Exercise 5.11. Assume a population of size 100 that expands to a size of 1000 in one generation. How many offspring does an individual have on average? Is there also a chance that one of the 100 individuals does not produce any offspring? What is the distribution of offspring numbers in this population?

Looking backward in time that means that two individuals at time $t+1$ choose the same ancestor, i.e. coalesce by time t with probability $1/N_t$. We used this kind of argument already in the computation of effective population sizes in Section 3. This already tells us a lot, e.g. in case the population size grew in the past then coalescence of the two lines happens faster as in each time step backward in time they have a higher chance of coalescence. In case there was a population size decline coalescence happens slower than expected.

Population expansion

The simplest model of a population size change is that of an exponentially growing (or declining) population. Assume a population of size N_0 colonizes a new habitat at time t_0 where it finds an abundance of ressources. Then the population can expand. How fast does it grow? That depends on many circumstances which we are not interested in from a modeling point of view. The simplest view is that on average an individual will leave a certain number of offspring (more than one for expansion and less than one for a decline). So the change in population size is proportional to the number of individuals currently living in the population. This leads to an exponential growth of the population. The population site at time t is given by

$$N(t) = \begin{cases} N_0, & t \leq t_0, \\ N_0 e^{\lambda(t-t_0)}, & t > t_0 \end{cases} \tag{5.1}$$

for some parameter λ which quantifies the speed of growth.

Graphically this looks like the upper curve in Figure 5.1. Furthermore in this figure you see an instance of a coalescent tree in this expanding population. Coalescence gets more probable the smaller the population size as the number of possible ancestors to choose from is smaller. So the coalescent tree in an expanding population looks squeezed compared to

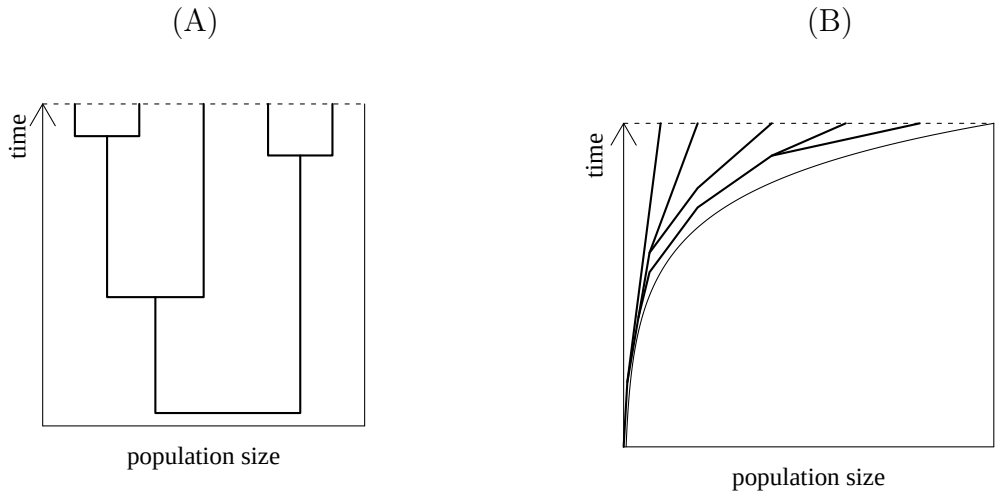


Figure 5.1: (A) The coalescent for a population of constant size and (B) for an expanding population.

a neutral one. Lineages stay uncoalesced a long time but when a first coalescence occurs it is very probable that the population size is already quite low. But when the population size is low the chance that more coalescences are to come soon is very high.

Comparing the coalescent trees in Figure 5.1 which come from an expanding population and a population of constant size we use population sizes such that the total tree lengths are approximately equal. For neutral mutations which afterwards can be seen in data that means that the number of neutral mutations on the trees are approximately equal.

Exercise 5.12. Consider the trees in Figure 5.1. Assume that on both trees 10 mutations have happened. As every generation and every line is as likely as any other to be hit by a mutation, these mutations are distributed randomly on the tree. Try to mimic a *random generator* and distribute the 10 mutations on both trees. (Observe that a uniform distribution does NOT mean that the mutations are all in approximately the same distances. A uniform distribution only means that any point on the tree is equally likely to be hit.)

Next, let us consider the effect of mutations falling on the tree on statistics like $\hat{\theta}_\pi$ and $\hat{\theta}_W$. Assume the MRCA of the sample has the sequence `AGTCTCGTGTTT`. Use the mutations you created to obtain the sequences that you would sample today, i.e. the sequences at the tips of the coalescent trees. For both the expanding and the constant population size calculate $\widehat{\theta}_S$ and $\widehat{\theta}_\pi$. Additionally draw the frequency spectrum in both cases. Did you really need the sequence of the MRCA to calculate $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$?

Exercise 5.13. Let us look at Figure 5.1. Assume the MRCA of the sample has the sequence `AGTCTCGTGTTT`. Use the mutations you created in the last exercise to obtain the sequences that you would sample today, i.e. the sequences at the tips of the coalescent trees. For both the expanding and the constant population size calculate $\widehat{\theta}_S$ and $\widehat{\theta}_\pi$. Additionally

draw the frequency spectrum in both cases. Did you really need the sequence of the MRCA to calculate $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$?

It is generally true that in the example with the expanding populations very few mutations will fall on the tree where coalescence has already happened for any two lines. That means that most mutations will only affect one individual in the sample. This is clearly seen in the frequency spectrum because here high frequency variants are missing. But it is also possible to make theoretical predictions how $\widehat{\theta}_\pi$ and $\widehat{\theta}_S$ differ for constant size and expanding populations. When both trees have the same length the number of mutations that fall on the tree will be similar and so $\widehat{\theta}_S$ is similar because it only uses the number of segregating sites. But $\widehat{\theta}_\pi$ will be different. As most mutations in the expanding population only affect one individual most mutations will only contribute to $n-1$ pairwise comparisons. This is much less than in the constant population size case. Here we have already calculated that among all mutations those affecting i individuals have an expected frequency proportional to $\frac{1}{i}$. So let Z be the number of individuals a randomly chosen mutation will affect. Then

$$\mathbf{P}[Z=i] = \frac{1}{a_{n-1}}\frac{1}{i}, \qquad a_m = \sum_{j=1}^{m}\frac{1}{j}.$$

Given a mutation affects i individuals it contributes to $i(n-i)$ different pairwise comparisons. So the expected number one mutation contributes to the pairwise comparisons is

$$\sum_{i=1}^{n-1} i(n-i)\mathbf{P}[Z=i] = \frac{1}{a_{n-1}}\sum_{i=1}^{n-1} n-i = \frac{1}{a_{n-1}}\sum_{i=1}^{n-1} i = \frac{(n-1)n}{2a_{n-1}}.$$

As $n > 2a_{n-1}$ this is bigger than $n-1$ which was the same number in the case of expanding populations. That means that compared with a neutral population of constant size $\widehat{\theta}_S$ can be the same but $\widehat{\theta}_\pi$ is much smaller in the case of an expanding population.

Exercise 5.14. The human population certainly did not have a constant size in the past. Can you somehow see this in the data from `TNFSF5`? In addition to that consider the European subpopulation and try to see population expansion here. In your analysis you use certain statistics that you (or `DNASP`) calculated from data. Which information do these statistics use from the data? E.g. the positions of the SNPs do not play a role for the value of most statistics, so the information stored in them is not used.

Are you confident with your findings for the European population?

Bottlenecks

Sometimes for a population the environment changes. This can lead to big challenges the population has to solve. Think e.g. of a new parasite entering the habitat of the population or a change in the climate. But before the population can recover from this change it reduces in size. After the change the population slowly grows again in size. This

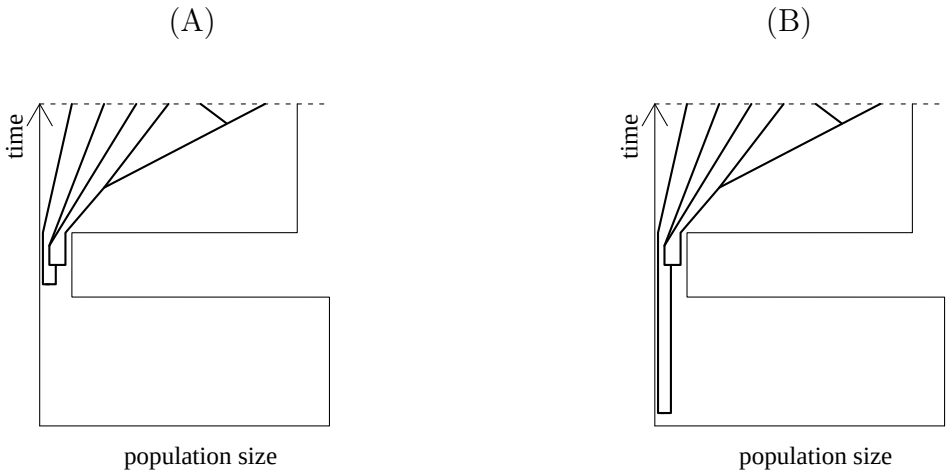


Figure 5.2: Two cases in a bottleneck mode. (A) Only one ancestral line survives the bottleneck. (B) Two or more lines survive which leads to different patterns in observed data.

scenario is known as a bottleneck. So here up to some time t_0 the population is of constant size N_0 and then it shrinks to some lower number. After that N_t grows again. We are considering the case when it grows again to its original size N_0.

Again individuals that are sampled today can be traced back in history and again the probability that two lines at time t find a common ancestor at time $t-1$ is $1/N_{t-1}$. This is seen in two examples in Figure 5.2. As for the expanding populations as there are few individuals at the time of the bottleneck shortly after the bottleneck there is a big chance of coalescence. In the above figure all lines coalesce. In that case the pattern observed in data looks exactly the same as for a population expansion. As there is no line leading before the time of the bottleneck there is no way of distinguishing the bottleneck model from the model of a population expansion. The effect is then also the same, reduction in high frequency variants and $\widehat{\theta}_\pi < \widehat{\theta}_S$.

However there is also the chance that two or more lines have not coalesced at the time of the bottleneck. As from then on backward in time the population is again larger coalescence will occur as in the case of a constant population size with that size. The effect in this case is exactly the opposite of the first case. As the lines surviving the bottleneck need some time to coalesce but already are ancestors of several individuals in the sample there will be an increase in high frequency variants.

But which of these two examples is more likely for a bottleneck? Well, that depends on several things. Above all, it depends on two things. First, on the length of the time when there is a bigger chance of coalescence. forward in time that is related to the rate at which the population recovers from the bottleneck. Second on the reduction of population size at time t_0 because the stronger the reduction the higher the chance of coalescence.

These two parameters, the rate how fast the population size recovers and the reduction in population size are usually combined to the *severity of the bottleneck* which is the product of these two parameters.

On the one hand one cannot be sure what happens under a bottleneck. On the other hand one can try to infer from data how strong the bottleneck really was. We do not go into details here.

5.4 The mismatch distribution

For two sequences we can calculate the probability that they are separated by k mutations. That is when in their coalescent k mutations on the two lines occur before they coalesce. As mutation rate is μ and coalescence rate is $1/2N$ and as there are two branches the probability that exactly one mutation before coalescence occurs is

$$\frac{2\mu}{2\mu + 1/2N} \cdot \frac{1/2N}{2\mu + 1/2N} = \frac{\theta}{\theta+1}\frac{1}{\theta+1}.$$

Here we have used a restart argument. Once the mutation event happened we can start the coalescent new and again the coalescence rate is $1/2N$ and the mutation rate for both lines is 2μ. This is true because of the Markoff property of the coalescent process. So the probability that there are k mutations separating the two lines is

$$\Big(\frac{\theta}{\theta+1}\Big)^k \frac{1}{\theta+1} \tag{5.2}$$

In a sample of size n there are $\binom{n}{2}$ pairs we compare. These pairs of DNA sequences were already used to calculate $\widehat{\theta}_\pi$ in (2.6). We can also ask for the distribution among pairwise differences, so e.g. how many pairs in our sample are separated by 5 mutations. This distribution is called the *mismatch distribution*. On the x-axis there is the number of mutations by which a pair is separated and on the y-axis the frequency of pairs with this number of differences.

What is the theoretical prediction for this mismatch distribution? In (5.2) we gave the probability that a pair is separated by k mutations. Approximately this formula should give the frequency of pairs with this number of differences.

Exercise 5.15. Draw the mismatch distribution for both samples from Exercise 5.12.

Assume $\theta = 0.4$. Draw the theoretical prediction from (5.2) into the figure of the last exercise.

So why is the theoretical prediction not met in this example? As simulations show it is seldomly met also for neutral populations of constant size. The statement that the frequency of *something* in a sample is close to the probability of *something* is often used but it relies on the independence of the sample. The pairs which build the sample in this case are highly dependent. This is for two reasons. First, as one individual appears in $n-1$ pairs these $n-1$ pairs are dependent. Second, all sampled individuals are part of *one* coalescent tree which makes them dependent.

Exercise 5.16. Assume a coalescent tree with three individuals. When the first two individuals are separated by 3 and the second and the third individual by 5 mutations there are only a few possibilities how the mutations fall on the coalescent tree. What is the minimal and maximal number of mutations by which the first individual is separated from the third?

This dependence of the sampled sequences is - from a statistical point of view - a major challenge in population genetics. It is also the reason why very many statistical standard techniques - e.g. t-tests - can rarely be used in population genetics.

However the mismatch distribution is still useful. Let us consider the expanding population example from Figure 5.1. There we saw that most coalescence events occur in the past when the population began to grow. That means that most pairs in the sample will coalesce around this time in the past. Given this time τ in generations the number of mutations that separate one pair is Poisson distributed with parameter $2\tau\mu$. Especially that means that we can expect most pairs to be separated by approximately $2\tau\mu$ mutations. For this reason in expanding populations we expect a peak in the mismatch distribution at around 2τ mutational units. Therefore this is a simple way to give a guess about the time of a population expansion.

Exercise 5.17. You already saw some datasets. Think of which populations should have signs of population expansions. Look at the data and try to see your conjectures in the data.

6 Recombination and linkage disequilibrium

In the beginning of the 20th century recombination was studied by the geneticist Morgan and his students. They were able to determine recombination rates between certain genes by doing crossing experiments and measuring genotype frequencies. Using these results they were able to infer genetic maps.

In this chapter, we will first describe the molecular basis of recombination. Then, we introduce recombination in teh Wright-Fisher model. We assume the recombination rate is already known (maybe from crossing experiments) and ask which theoretical predictions come from models with recombination, and which consequences do these predictions have for data taken from a sample of a population. One consequence, linkage disequilibrium, is discussed In the last part of the chapter.

6.1 Molecular basis of recombination

In diploid organisms recombination happens during meiosis. Recombination mixes parental and maternal material before it is given to the next generation. Each gamete that is produced by an individual therefore contains material from the maternal and the paternal side. To see what this means, let us look at your two chromosomes number 1, one of which came from your father and one from your mother. The one that you got from your father is in fact a mosaic of pieces from his mother and his father, your two paternal grandparents. In humans these mosaics are such that a chromosome is made of a couple of chunks, more than one, but probably less than ten. Chromosomes that don't recombine are not mosaics. The Y-chromosome doesn't recombine at all, you get it completely from your father and your paternal grandfather. Mitochondrial DNA also doesn't normally recombine, (although there is evidence for some recombination, see Eyre-Walker *et al.* (1999)), you normally get the whole mitochondria from your maternal grandmother. The X-chromosome only recombines when it is in a female.

Exercise 6.1. Who, of your 4 grandparents contributed to your two X chromosomes (if you are female) or single X chromosome (if you are male)?

Figure 6.1 shows how parts of chromosomes are exchanged. The picture shows what will become 4 gametes. Here you see the effect of *crossing over*, which is the most well-known recombination mechanism. Parts of homologous chromosomes are exchanged. Crossing over only happens in diploid individuals. However, exchange of genetic material can also happen in haploid individuals. In this case two different individuals exchange pieces of their genome. In a diploid, recombination only makes a difference if the individual is not completely homozygous. *Gene conversion*, which is another mechanism of recombination, will not be treated here. We will therefore use the term recombination as a synonym for crossing over.

Mendels second law (independent assortment) states that genes are inherited independently of eachother. This is in fact only true for genes that lie on different chromosomes or that are far away from each other on the same chromosome. It means that the probability

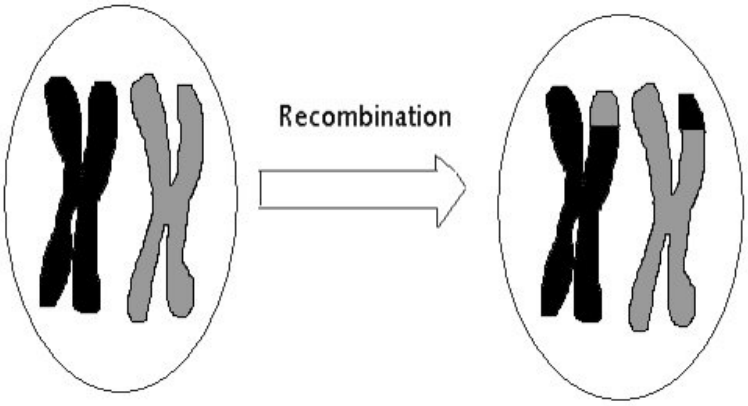


Figure 6.1: Recombination (Crossing Over) of chromosomes.

of inheriting an allele from your grandmother at one chromosome doesn't tell you anything about the probability that you also inherit other alleles from your grandmother located on different chromosomes. On the other hand, if genes are on the same chromosome they are said to be *physically* linked, if genes are very close to each another recombination between them is very rare and they will be inherited together. So when considering two loci, the two extreme cases are: completely unlinked genes leading to independent inheritance and completely linked genes that are always inherited together.

Let us first look at an example of independent inheritance, that is a case where Mendels second law is correct. We are looking at two genes that are on different chromosomes. Mendel mated a plant that was homozygous for round R and yellow Y seeds with a plant that was homozygous for wrinkled r and green y seeds. R and Y are dominant alleles (indicated by capital letters here) which means that the phenotype of the plant is determined by the presence or absence of this allele no matter what the second allele is. All F_1 offspring are $Rr\,Y\,y$ (which is called a *dihybrid*) and they all have round and yellow seeds. In controlled crosses the parent generation is referred to as the P generation, the offspring of the P generation is called the F_1 generation, the generation after that the F_2 generation etc. To simplify the analysis, we (virtually) mate the dihybrids $Rr\,Y\,y$ from the F_1 generation with a homozygous recessive strain $rr\,yy$. Such a mating is called a test cross because it exposes the genotype of all the gametes of the strain being evaluated. Because of the independence of inheritance, the probability of finding a $Rr\,yy$ individual will be the same as the probability of finding a $Rr\,Y\,y$ individual.

Exercise 6.2. Calculate the probabilities for the test-cross offspring for each possible genotype. And calculate the probabilities for the phenotypes.

Linked genes in a testcross

Now we will treat the case intermediate between the two extremes, i.e. partially linked genes. Imagine that the two genes would have been on the same chromosome and close to each other. Remember that the individuals that Mendel started with (the P-generation) were double homozygotes, so that if they had a Y they certainly also had an R. This Y and R would stay together if they were close to each other on the same chromosome.

We start with two different strains of corn (maize). One that is homozygous for two traits, yellow *colored* kernels CC which are filled with endosperm causing the kernels to be *smooth* $ShSh$ and a second that is homozygous for *colorless* kernels cc that are wrinkled because their endosperm is *shrunken* $shsh$.

When the pollen of the first strain is dusted on the silks of the second (or vice versa), the kernels produced (in the F_1 generation) are all yellow and smooth. So the alleles for yellow color C and smoothness Sh are dominant over those for colorlessness c and shrunken endosperm sh. Again we do a testcross (mate the $F1$ with recessive double homozygotes) because it exposes the genotype of all the gametes of the strain being evaluated. In this example the genes are so close that only 2.8% of the offspring is $cc\,Shsh$ or $Cc\,shsh$. All the others are $cc\,shsh$ or $Cc\,Shsh$. Instead of 25% we find 48.6% of the genotype $cc\,shsh$.

Note that recombination rate can be given per nucleotide or as the probability that a recombination event happens between two loci on a chromosome. Generally, in population genetics, it will be given as the per nucleotide per generation recombination rate.

Exercise 6.3. 1. In the last example the probability of recombination was 2.8% per generation. If this probability would be close to 0, what would be our conclusion about the location of these genes?

2. Given that the per nucleotide per generation recombination rate is 10^{-8} how far away from each other do you think the genes for C or c and Sh or sh are?

6.2 Modeling recombination

We will look at the Wright-Fisher model again. In this model we imagined that offspring chooses a single parent at random. When we include the possibility of recombination, it is maybe more natural to think about single chromosomes having parent chromosomes. So you can think of two loci on your chromosome 1, we call them locus A and B. We only look at one copy of your chromosome 1, the one you got from your mother. Now, if we trace back the history of the loci A and B, the step back to your mother is obvious,but then in the next step, they could have come from her mother or from her father. We first decide for locus A where it came from - let's say it came from your grandfather. Now with a certain probability, the allele at locus B came from your grandfather too (if there was no recombination) and with a certain probability from your grandmother (if there was recombination).

Forward in time let the probability that a recombination event occurs be ρ. Then

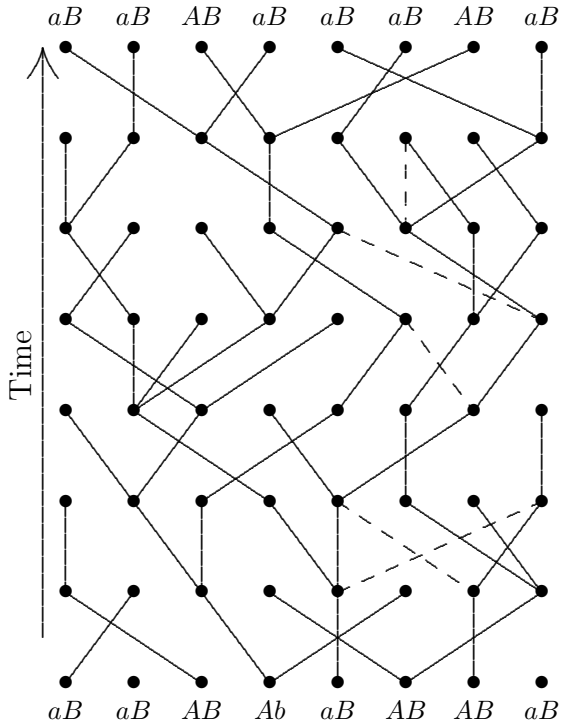


Figure 6.2: The Wright-Fisher model with recombination (see text). At certain points a recombination event happens (dashed lines). This leads to a different ancestry of the A/a and the B/b locus. The B/b-locus is inherited via the dashed lines whereas the A/a locus goes along with the solid lines.

backward in time the probability of choosing the same parent is

$$\mathbf{P}[\text{two loci have the same parent}] \approx 1 - \rho.$$

The approximation here is that no more than one recombination event occurs between the two loci which has a probability of order ρ^2 and may be neglected. If a (one) recombination event happened between the two loci, the second locus will have a different parent. As long as ρ is small this is the probability of choosing a different parent. In other words, with probability $1-\rho$ the second locus is on the same chunk of DNA and it will be derived from the same grandparent (grandmother or grandfather).

Taking reality not too literally we can say that the two ancestors are chosen randomly in the whole population. This is not true for many reasons, one being that when the two chunks of DNA choose two different parents necessarily they come from individuals of opposite sexes. But one generation before that they come from a male and a female with probabilities proportional to the males and female frequency in the population. Figure 6.2 shows a cartoon of the Wright-Fisher model with recombination.

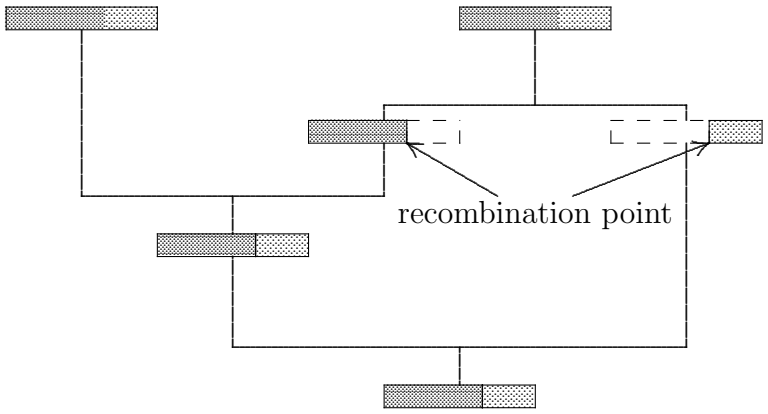


Figure 6.3: The coalescent with recombination starting with a sample of size 2. One recombination event splits the chromosome in two parts. The first shares its left part with the sample, the second one the right part.

Recombination and the coalescent

The coalescent was used as a stochastic process that gives the ancestry of a sample of sequences. However, just like in the Wright-Fisher model, if we include recombination, there is a chance that two loci on a chromosome do not have the same ancestry. How can this be built into the coalescent?

Take one lineage. Tracing it back recombination events can happen. Recombination occurs with probability ρ each generation, therefore the waiting time until the first recombination event is geometrical with success parameter ρ. This is almost the same as an exponential with rate ρ (see Maths 2.4). If we look at more than one lineage, both recombination and coalescence can happen. As we can see from the Wright-Fisher model two lines find a common ancestor with probability $1/2N$ per generation. And recombination happens with probability ρ in every lineage in every generation. Even though there are now more events possible, we still assume that not more than one event happens at any given time. We now have as an approximate process the following:

- Given n lines coalescence of two of them occurs with rate $\frac{\binom{n}{2}}{2N}$. After coalescence $n-1$ lines are present in the coalescent process.

- A recombination event occurs in each line with rate ρ. Given that a line in the coalescent tree shares both loci with the sample these loci split leading to a split of one line into two lines one carrying locus A and the other carrying locus B.

This process is illustrated in a very simple case in Figure 6.3.

Many coalescent trees for a stretch of DNA

As one moves along the genome, one will have a series of sites that all have the same coalescent tree, with no recombination anywhere in that tree. But eventually one hits a site where the genealogy changed. This will cause a particular kind of rearrangement in the genealogy. That tree will then hold for a while as one moves along the genome and then there will be another breakage and re-attachment. After some distance the tree has changed to be totally different.

But starting at some site how far will we have to go until the next recombination event? You might think that this is far apart but it is not. In humans one expects about one recombination event every 10^8 bases every generation. So $r = 10^{-8}$ (we refer to the recombination rate between two adjacent nucleotides with r and between more distant loci with ρ) in the coalescent with recombination. Given two sites that are d base pairs apart, the recombination rate is $\rho = rd$. So, if we consider two lines coalesce with rate $1/2N$ and recombination with rd we are interested when to expect a recombination between the two sites. Let us say we want to calculate the distance we need in order to have a 50% chance to have a recombination event between the site. Then we have to solve

$$\mathbf{P}[\text{recombination before coalescence}] = \frac{2rd}{2rd + 1/2N} = 1 - \frac{1}{4Nrd + 1} \geq 0.5$$

for d. Again we used the competing exponentials from Maths 3.1. This gives

$$4Nrd \geq 1, \qquad d \geq \frac{1}{4Nr}.$$

For humans we have $N_e \approx 10^4$ and $r \approx 10^{-8}$. This gives

$$d \geq \frac{1}{4N_e r} \approx 2500.$$

If the human population size would be 10^5 this number changes to only 250 bases. In Drosophila, where the effective population size is larger the distance is much shorter (about 10 to 100 times shorter). PSP: shouldnt we do this for a sample of size n or 10 orso instead of 2?

We must think of the coalescent trees in a genome as each holding for only a small region in the genome, and that there may be a million different coalescent trees that characterize the ancestry of our genome. Virtually every locus has its own most recent common ancestor (MRCA), at widely differing times and places. As put in Felsenstein (2004): "We not only have as our ancestors mitochondrial Eve and Y-chromosome Adam (who did not know each other) but also hemoglobin Sam and cytochrome Frieda, and a great many others as well."

The number of lineages in a coalescent tree with recombiation

Let us make one more example for the coalescent with recombination. For a tree of a sample of 5, going back in time two things can happen: a coalescent event or a recombination

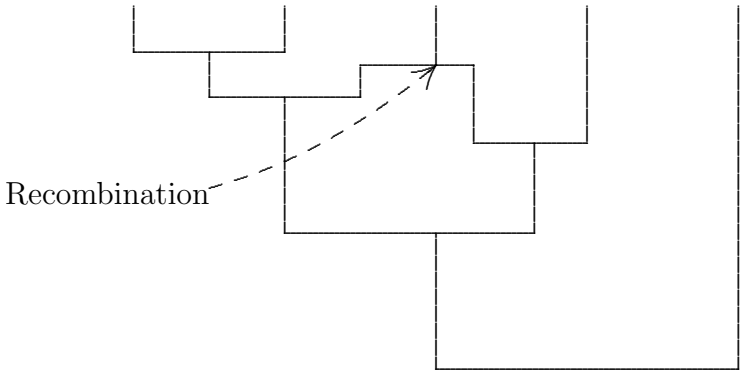


Figure 6.4: The coalescent with recombination for a sample of 5 individuals. One recombination event leads to different ancestries at two loci on the chromosome.

event. Recombination occurs at rate rd and when there are currently n lines, coalescence occurs at rate $\frac{\binom{n}{2}}{2N}$. So when the process starts this rate is $\frac{10}{2N}$. Suppose first a coalescence event happens (the number of lineages goes to 4) and then a recombination event happens (the lineage splits and there are then again 5 branches) and then only coalescence events happen. In Figure 6.4 you find a picture. Because recombination events increase the number of lineages it can take a long time before all lineages find a common ancestor. However, each time that the number of lineages grows, the coalescence rate increases, whereas the recombination rate stays the same. Therefore in finite time the sample will find a common ancestor.

Exercise 6.4. Look at Figure 6.4. Let us assume in the recombination event the left part of the chromosome takes the left branch and the right part of the chromosome takes the right branch. Draw the genealogy of the left branch and the right branch in two separate trees.

The importance of recombination

In humans and Drosophila mutation and recombination tend to happen on the same scale, that means if you expect to find mutations on a tree you can also expect to find recombinations. This is just a coincidence because μ and r in humans are both on the order of 10^{-8}.

Recombination is very important for population genetics research. Without recombination all loci on a chromosome would share one tree and they would basically constitute one data-point. We would never be able to get more data points than there are chromosomes. Fortunately, recombination makes loci that are relatively far away from each other on the same chromosome independent, thereby dramatically increasing the amount of information that we can extract from sequence data.

Exercise 6.5. Sequencing costs time and money. Suppose you have data for 10 individuals for one locus and the data look like there was a population bottleneck in the past. However, the reviewers of your manuscript don't believe it yet and ask for more data. Now, you could either

- sequence a larger sample at the same locus,
- sequence several loci in the same sample.

What would you suggest and why?

6.3 Recombination and data

The four-gamete-rule

Recombination affects the genealogical tree of the sequenced sample. So it must also affect the SNP data we obtain by sequencing. The *four-gamete* rule is a handy rule to see quickly whether recombination *must* have happened in the history of a sample. The main idea is that recombination will make two parts of a sequence have different trees. Look again at Figure 6.4 and recall Exercise 6.4. In the tree of the left part of the chromosome, there could be mutations that affect sequence 1, 2 and 3 at the same time. Imagine that there is a nucleotide where a mutation took place from a `C` to a `G` affecting individuals 1, 2 and 3. 1, 2 and 3 will now carry the `G` there and the others (4 and 5) a `C`. We could write this in a table as `DNASP` does:

```
1  G
2  G
3  G
4  C
5  C
```

Other mutations on the tree could affect 1 and 2, or 1, 2, 3 and 4 or a mutation can affect just one of the individuals. If it had affected 1 and 2, it would show in the table as follows:

```
1  G  A
2  G  A
3  G  T
4  C  T
5  C  T
```

Exercise 6.6. Now draw an unrooted tree of the 5 individuals and indicate where the mutations must have happened.

Given the tree that you have drawn, it is impossible that a mutation affects 3 and 4, but not 5, because there is no branch on the tree that is shared by 3 and 4 and not 5. However, if recombination changes the tree, than suddenly other combinations of individuals share branches and mutations can affect them. In Figure 6.4, the right tree allows for 3 and 4 to share a mutation that is not shared with 5. In the table this would look like this. And from the information in the table, you can see that recombination must have taken place.

1	G	A	C
2	G	A	C
3	G	T	A
4	C	T	A
5	C	T	C

To immediately see that recombination must have taken place you don't need to draw trees, you can use the *four-gamete-rule*:

> If you can find four different gametes (which is the same as genotype or haplotype) in a sample, by considering just two (diallelic) polymorphic sites, a recombination event must have taken place between the two sites.

In the example, if you look at the first and third mutation, you find that 1 and 2 have genotype `GC`, 3 has `GA`, 4 has `CA` and 5 has `CC`, which are four different genotypes. You can now immediately conclude that recombination must have taken place between the two sites.

Exercise 6.7. Can you tell from the last table to decide whether recombination has taken place to the left or to the right of the second mutation?

Exercise 6.8. The aim of this exercise is to find out the number of recombination events in the stretch of DNA that is sequenced in the `TNFSF5` study. The sequences are about 5000 nucleotides long.

1. How many recombination events do you expect to find? To calculate this compute the average tree length (or find it in a different Section of this manuscript) and the average number of recombination events given this length.

2. Open in `DNASP` the file `TNFSF5.nex`. First of all you only need the different haplotypes in the sample because the four-gamete-rule works with haplotypes. Use `Generate->Haplotype Data File` to produce a file only consisting of the different haplotypes. Do you see if recombination has taken place in the past?

3. You can also ask `DNASP` to look for recombination (use `Analysis->Recombination`). How many pairs of segregating sites do you find where the four-gamete-rule detects recombination? What is the minimum number of recombination events in the history of the sample?

Exercise 6.9. Recombination has some impacts on sequence variation. Let us produce one of them using `seqEvoNeutral()` from the `R`-package.

Compare the evolution of the site frequency spectrum for sequence evolution with and without recombination. Use `N=25, u=1, stoptime=200, seq=FALSE, sfs=TRUE`. Do one simulation with `r=0` and one with `r=1`. Which of the two final plots is closer to the neutral expectation? Can you explain your result?

Linkage Disequilibrium and recombination

We argued that linkage between loci can be broken up by recombination. The amount to which loci are linked can be measured using *linkage disequilibrium*. Let us again consider a model with two loci both of which have two alleles. All combinations of alleles are found in Figure 6.5. The probability that a recombinant gamete is produced at meiosis was denoted by ρ. A different measure is the *genetic map distance* of two loci which is always greater that ρ because it is the average number of recombinational events rather than the probability of producing a recombinant offspring.

Figure 6.5 shows that there are four gametes in the population, A_1B_1, A_1B_2, A_2B_1, and A_2B_2 with frequencies p_{11}, p_{12}, p_{21} and p_{22} respectively. The frequency of the A_1 allele as a function of the gamete frequencies is $p_{1\cdot} = p_{11} + p_{12}$. Similarly the allele frequency of B_1 is $p_{\cdot 1} = p_{11} + p_{21}$. Recombination changes the frequencies of the 4 gametes. For example, the frequency of the A_1B_1 gamete after a round of random mating, p'_{11} (where the $'$ stands for the next generation) is

$$p'_{11} = (1-\rho)p_{11} + \rho p_{1\cdot}p_{\cdot 1}. \tag{6.1}$$

This expression is best understood as a statement about the probability of choosing an A_1B_1 gamete from the population. We use the law of total probabilities again. A randomly chosen gamete will have had one of two possible histories: either it will be a recombinant gamete (this occurs with probability ρ) or it won't (with probability $1-\rho$). If it is not a recombinant, then the probability that it is an A_1B_1 gamete is p_1. Thus, the probability that the chosen gamete is an unrecombined A_1B_1 gamete is $(1-\rho)x_1$ which is the first term on the right side of (6.1). If the gamete is a recombinant, then the probability that it is a A_1B_1 gamete is the probability that the A locus is A_1, which is just the frequency of A_1 which we denote by $p_{1\cdot}$, multiplied by the probability that the B locus is B_1, $p_{\cdot 1}$. The probability of being a recombinant gamete and being A_1B_1 is $\rho p_{1\cdot}p_{\cdot 1}$. When it is a recombinant the two loci are chosen independently and so we have to multiply frequencies of A_1 and B_1 which are $p_{1\cdot}$ and $p_{\cdot 1}$.

Exercise 6.10. Derive the three equations for the frequencies of the A_1B_2, A_2B_1, A_2B_2 gametes after a round of random mating.

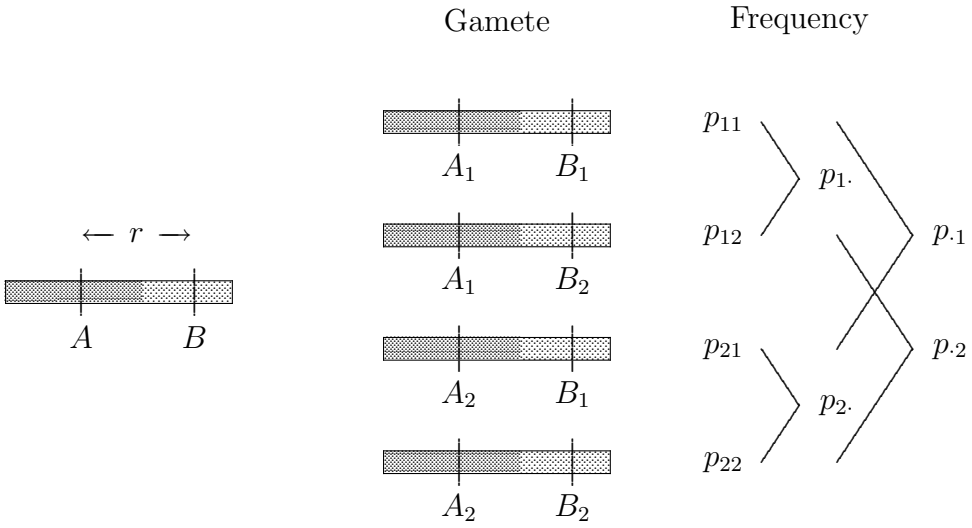


Figure 6.5: In the model with two loci and two alleles per locus four different gametes are possible.

The change in the frequency of the A_1B_1 gamete in a single generation of random mating is, from (6.1)

$$\Delta_\rho p_{11} = -\rho(p_{11} - p_{1\cdot}p_{\cdot 1}) \tag{6.2}$$

where Δ_ρ measures the change in LD due to recombination. Linkage means that the loci do not behave independently. We can define a linkage disequilibrium parameter D as

$$D = p_{11} - p_{1\cdot}p_{\cdot 1}, \tag{6.3}$$

which is the difference between the frequency of the A_1B_1 gamete, p_{11}, and the expected frequency if alleles associated randomly on chromosomes, $p_{1\cdot}p_{\cdot 1}$. We can now write

$$\Delta_\rho p_{11} = -\rho D.$$

The equilibrium gamete frequency is obtained by solving

$$\Delta_\rho p_{11} = 0 \qquad \text{and so} \qquad p^*_{11} = p_{1\cdot}p_{\cdot 1}.$$

So we have concluded formally what we already knew intuitively: recombination reduces LD. The time scale of change of gamete frequencies due to recombination is roughly the reciprocal of the recombination rate. (PSP: dont understand last sentence)

The frequency of the A_1B_1 gamete may be written

$$p_{11} = p_{1\cdot}p_{\cdot 1} + D,$$

which emphasizes that the departure of the gamete frequency from its equilibrium value is determined by D. We can rearrange the terms of the above definition of D to obtain

$$\begin{aligned} D &= p_{11} - p_{1\cdot}p_{\cdot 1} = p_{11} - (p_{11} + p_{12})(p_{11} + p_{21}) \\ &= p_{11} - p_{11}(p_{11} + p_{12} + p_{21}) - p_{12}p_{21} = p_{11} - p_{11}(1 - p_{22}) - p_{12}p_{21} \\ &= p_{11}p_{22} - p_{12}p_{21}. \end{aligned} \tag{6.4}$$

Calculating the same things for the other associations between A_1, A_2 and B_1, B_2 we obtain the following:

Gamete:	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency:	p_{11}	p_{12}	p_{21}	p_{22}
Frequency:	$p_{1\cdot}p_{\cdot 1} + D$	$p_{1\cdot}p_{\cdot 2} - D$	$p_{2\cdot}p_{\cdot 1} - D$	$p_{2\cdot}p_{\cdot 2} + D$

Exercise 6.11. Show that the gamete frequencies as a function of D are correct in the above table.

The A_1B_1 and A_2B_2 gametes are sometimes called coupling gametes because the same subscript is used for both alleles. The A_1B_2 and A_2B_1 gametes are called repulsion gametes. Linkage disequilibrium may be thought of as a measure of the excess of coupling over repulsion gametes. When D is positive, there are more coupling gametes than expected at equilibrium; when negative, there are more repulsion gametes than expected.

Using linkage disequilibrium we want to measure association between alleles. In statistical terms association results in the correlation of certain random variables. The measure D can also be seen in this way as a covariance of two random variables. (See Maths **??**.) But which random variables must we take in order to see D as the covariance. Let us pick (virtually) one individual from the population and set

$$\begin{aligned} X &= \begin{cases} 1, & \text{if we find allele } A_1, \\ 0, & \text{if we find allele } A_2, \end{cases} \\ Y &= \begin{cases} 1, & \text{if we find allele } B_1, \\ 0, & \text{if we find allele } B_2. \end{cases} \end{aligned} \tag{6.5}$$

Then

$$\mathbf{E}[X] = \mathbf{P}[X = 1] = p_{1\cdot} \quad \text{and } \mathbf{E}[Y] = \mathbf{P}[Y = 1] = p_{\cdot 1}.$$

Then their covariance is

$$\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X] \cdot \mathbf{E}[Y] = p_{11} - p_{1\cdot}p_{\cdot 1} = D.$$

Now let us see how D evolves. The value of D after a round of random mating may be obtained directly from equation (6.3) by using $p_{11} = p_{1\cdot}p_{\cdot 1} + D$ to be

$$\begin{aligned} D' &= p'_{11} - p'_{1\cdot}p'_{\cdot 1} = (1-\rho)p_{11} + \rho p_{1\cdot}p_{\cdot 1} - p'_{1\cdot}p'_{\cdot 1} \\ &= (1-\rho)(p_{11} - p_{1\cdot}p_{\cdot 1}) = (1-\rho)D \end{aligned} \tag{6.6}$$

Here we have used that $p'_{1\cdot} = p_{1\cdot}$ and $p'_{\cdot 1} = p_{\cdot 1}$ which is approximately true at least in large populations as drift can be neglected there. The change in D in a single generation is thus

$$\Delta_\rho D = -\rho D,$$

which depends on the gamete frequencies only through their contributions to D. Finally this equation gives nothing but a geometric decay of D, meaning that

$$D_t = (1-\rho)^t D_0,$$

showing, once again, that the ultimate state of the population is $D = 0$.

Exercise 6.12. Consider two loci A and B with $\rho = 0.5$. In this case in the diploid organism the probability of inheriting the A locus from one individual is independent of whether or not the B locus is inherited from this chromosome. This is the reason why one speaks of $\rho = 0.5$ as *free recombination*. However, linkage disequilibrium measured by D

does not vanish immediately after one generation of random mating. Can you clarify why it doesn't? Did we do something wrong in our calculations? (Tip: write down for two generations the genotype freqs and the gamete freqs when you start with half of the pop $AABB$ and half $aabb$)

In natural populations, the reduction in the magnitude of linkage disequilibrium by recombination is opposed by other evolutionary forces that may increase $|D|$. We will come to some of them in the next Section.

We saw that D can be seen as the covariance of two random variables. Unfortunately D is very sensitive to allele frequencies. As frequencies must be positive we can read from the table of allele frequencies that, when D is positive,

$$D \leq min\{p_1.p_{\cdot 2}, p_2.p_{\cdot 1}\}.$$

A normalized measure taking account of this and which is therefore much less sensitive to allele frequencies is

$$r^2 = \frac{D^2}{p_1.p_{\cdot 1}p_2.p_{\cdot 2}}.$$

Again this can be seen as a statistical measure.

Maths 6.1. *Given the random variables X and Y of Maths* **??** *define the* correlation coefficient r_{XY} *of X and Y as*

$$r_{XY} = \frac{|\mathbf{Cov}[X,Y]|}{\sqrt{\mathbf{Var}[X]\mathbf{Var}[Y]}}.$$

So taking the random variables from (6.5) we have, as X and Y are binomial with success probabilities $p_1.$ and $p_{\cdot 1}$ respectively that r^2 is the square of the correlation coefficient of these two random variables.

Exercise 6.13. `DNASP` calculates D and r^2. Use the `TNFSF5` data set to see some values.

1. Our above calculations were done in a model. Now `DNASP` has calculated something from real data. Is D in the above equations really the same D as the D that is calculated by `DNASP`?

2. Linkage should decay with distance between two segregating sites. `DNASP` is capable of analysis that addresses this question. Can you find out what the program does?

3. Two locus data can be arranged in a contingency table as you saw on page 57 (this time with only two columns). Whenever data can be arranged like this, a χ^2 test can be done. And, additionally, if it is a 2×2 contingency table, Fishers exact test can be used, which is a better test in this case which can also be computed by `DNASP`. Using `DNASP`, is there a significant amount of linkage disequilibrium in the dataset?

6.4 Example: Linkage Disequilibrium due to admixture

Since linkage disequilibrium vanishes exponentially fast, we must ask why it can still be observed in real data. One example is the situation of admixture of two populations. A population which recently was separated in two different populations does not need to be in linkage equilibrium.

Exercise 6.14. We will simulate the situation of an admixed population using `linkage()` from the `R`-package. Before you start read `?linkage` and make yourself familiar with details of the model.

1. Assume the loci A and B are on different chromosomes. What is the rate of recombination between them? Set the parameters on the model to represent the following scenario:

   - A and B loci on different chromosomes
   - `pop1` was initially fixed for the A_1 allele and the B_2 allele
   - `pop2` was initially fixed for the A_2 allele and the B_1 allele

   Create a graph that displays gamete/haplotype frequency $(p_{11}, p_{12}, p_{21}, p_{22})$ over time.

2. What is the frequency of each haplotype in the first generation? (You can use `ret<-linkage(...)` and `ret$Haplotype.freq[1,]`.) Explain why there are no individuals with the A_1B_1 or A_2B_2 haplotypes in the population.

3. Calculate D for the population in the first generation and check using `ret$LD[1]`. Which haplotypes are more common than you would expect if the population were in linkage equilibrium? Which are less common?

4. Assume that both populations are large such that genetic drift does not operate. Judging from the graph, about how many generations does it take for the gamete frequencies in the mixed population to stop changing? What is the frequency of each haplotype at that point? Using these haplotype frequencies, calculate the value of D.

5. What does this value of D say about gamete frequencies in the population (i.e. what does linkage equilibrium mean)?

Exercise 6.15. We have encountered Hardy-Weinberg and linkage equilibrium. Moreover, both equilibria can interact with genetic drift in small populations which we will investigate next.

1. Take the same initial frequencies as in Exercise 6.14. Compute by hand the inbreeding coefficient at the A and B locus at the time of the admixture. In how many generations is Hardy-Weinberg equilibrium reached?

2. Use `n.pop1=100`, `n.pop2=100`, set `stoptime=200` and vary `r` from 0.001 to 0.1. Plot D by using `what="LD"`. For $r = 0.001$ you sometimes do not see a decay in linkage disequilibrium at all, somtimes it goes down to 0. Can you explain this? In contrast, do you see a clear decay in D for $r = 0.1$?

Exercise 6.16. In Exercise 6.13 you calculated if the human population was in linkage equilibrium or not. But is the human population in Hardy-Weinberg equilibrium at the `TNFSF5` locus?

Exercise 6.17. A survey of human blood group and serum protein frequencies in a Michigan town revealed that the loci controlling these neutral phenotypes were in Hardy-Weinberg equilibrium (i.e. genotypes at each locus were observed at their expected Hardy-Weinbarg frequencies). Does this information tell you anything about the history of the population more than 1 generation ago? Explain why or why not.

The data also showed that the gamete frequencies were in linkage disequilibrium. What does this suggest about the length of time that the population has been mating at random? If you know that the area of study was settled recently by people from different parts of the world, what hypotheses might you form about the cause of linkage disequilibrium at these loci?

7 Various forms of Selection

The neutral theory, which was mainly developed by Motoo Kimura since the 1950s, assumes that every individual in a population has the same chance to produce offspring. This contradicts the Darwinian view on evolution that fitter individuals will produce more, or more viable offspring and eventually form the basis of all future generations. Which view is more realistic cannot be said generally. A big step towards unifying these two approaches is to attribute selection coefficients to certain genotypes. Roughly, the selection coefficient of an allele tells you how much more (or less) offspring the carrier of the allele is expected to have, compared to a reference allele. In a population of constant size, the expected number of offspring is 1 for every individual, if a new allele in the population has selection coefficient 0.05 this means the expected number of offspring of the carrier of the allele is 1.05. In this case the allele is said to have a fitness effect, and selection can only act if there are fitness differences caused by alleles with fitness effects. In Kimura's view, selection coefficients are usually close to zero, because most mutations are neutral, in other words, have no fitness effect.

Originally fitness is seen as a trait of a phenotype. Remember that high fitness means that an individual produces much viable offspring that contributes to future generations. As the phenotype has its genetic basis, we can also attribute fitness to a genotype. However, the question how genotypes translate to phenotypes is a very difficult one. In population genetic models we therefore attribute fitness directly to genotypes. A consequence of this is that it conceals the mechanisms where these fitness differences really come from.

7.1 Selection Pressures

What forces are responsible for differences in fitness of certain genotypes? Many mechanisms have been suggested.

The simplest example of an allele that has a fitness effect (this means it has a selection coefficient that is not zero) is if the allele produces a different protein (which is of course only possible in coding regions) which then can have positive or negative effects on the chances to produce viable offspring. Prominent examples are mutations in insects that results in insecticide resistance or, rather the opposite, mutations that cause an essential protein not to function anymore, so that the individual dies. But mutations that don't change proteins can also have fitness effects, for example if the mutation changes the abundance of a protein.

Selection can act on different parts of the life cycle of an individual. Let us give some examples:

Viability selection

Individuals with a high fitness not only produce much offspring but also produce viable offspring that reach maturity to produce offspring itself. Growing to adulthood is a big challenge for an individual. At first it has to survive as a zygote which can already be

hard. (Also in humans natural abortions are very common.) And then it has to survive until adulthood.

Sexual selection

Having evolutionary success means to have a large number of offspring, but in sexual species this is only possible if you have a mate. Some phenotypes have a higher chance to find a mate. This can e.g. be due to assortative mating which means that individuals prefer to mate with individuals that are alike. Assortative mating is common (e.g. in human populations where mating occurs according to social status), but disassortative mating also occurs (e.g. in plants where certain alleles can only reproduce when they mate with a different allele. These are called *self-incompatibility alleles*). Sexual selection can also be due to male competition and/or female choosiness. This is well-known from peacock-males and fights for the best female in bighorn sheep.

Gametic selection

Gametes themselves can also be more or less successful as they can also be more or less viable, e.g. because the genotype determines the protein constitution of a sperm cell. In meiosis it is possible that certain genotypes are more likely to produce gametes than others. This is known as *meiotic drive* which results in a non-Mendelian segregation of the alleles. One can recognize meiotic drive if an excess of gametes of a heterozygous individual carry the same genotype.

Fecundity selection

Fitness also depends on how many gametes an individual produces. This is known as *fecundity* (or *fertility*) *selection*. In male plants fitness depends on how many seeds the individuals pollen can pollinate, if it produces more pollen it will pollinate more seeds. For female plants the individual will have more offspring if it produces more seeds. In animals that take care of their young fertility selection acts on the family size.

The above list gives examples of selective forces that do not depend on the other individuals in the population. This is not always the case. The most important examples are

Density and frequency dependent selection

Density dependent selection means that fitness of an individual depends on the density of the population. This happens if certain phenotypes do well when the population density is low whereas other phenotypes do well when there is a lot of competition due to a high population density. This is called *density-dependent selection*.

Often, the frequency of a genotype in the population plays a role. Think e.g. of an allele that can only reproduce if it mates with a different allele, such as the self-incompatibility

allele of a plant. For those alleles it is clearly an advantage when they are in low frequency in the population because the chance to produce offspring is higher. Their fitness will go down if their frequency goes up. This is called negative *frequency-dependent selection*.

Exercise 7.1. Can you think of an example of positive frequency dependence? Do you think you are likely to find such examples in nature? □

Several more complicated effects that have to do with selection have been described. We mention two of them.

Pleiotropy

It is possible that the alleles at one locus affect several phenotypic traits of an individual. This is called pleitropy. A gene affecting the embryonic growth rate may also affect the age of reproduction of the individual. These traits can either be both selectively advantageous or can point to different directions. If a genotype increases the fertility of an individual but reduces its viability the overall effect on fitness may be small.

Epistasis

Many quantitative traits, e.g. height or weight, are affected by many genes. These genes can have effects that depend on other genes. The simplest example is when at one locus, the gene produces a protein if the individual carries allele A, and doesn't produce anything if the individual carries allele a. At another locus, the gene determines the abundance of the protein, it has two alleles, B for producing a lot of protein, b for producing half the amount. The effect of the alleles at the second locus now depends on the allele at the first locus. An individual carrying an A will produce more protein if it carries a B than it would if it would carry a b. However, an individual carrying an a will produce nothing, independent of whether it carries a B or a b. The dependence of one trait on different loci is called epistasis. Epistatic selection is complicated and will be ignored in the rest of the course.

Exercise 7.2. Various forms of selection were introduced above. From your biological experience find five examples where selection seems obvious to you. □

7.2 Modeling selection

The main aim of mathematical modeling is to find simple models that capture most (or all) of the biological features. These models can then be used for several things, for example to make predictions. Population geneticists are mainly interested in changes in allele and genotype frequencies, therefore population genetic models should take the forces into account that are responsible for these changes. In principle we would have to model the following:

	meiosis		*random union*		*survival*	
Adults	$\Longrightarrow$	Gametes	$\Longrightarrow$	Zygotes	$\Longrightarrow$	Adults
(N)	*fertility selection*	(∞)	*sexual selection*	(∞)	*viability selection*	(N)

To start with simple models for this scenario we do not consider the kind of selection too closely. In fact, we will only deal with viability selection here (as is usual in population genetics). The population consists of adults and has discrete generations. Consider two homologous alleles, A_1 and A_2 that are present in a population. If an individual has genotype A_1A_1 then its probability of surviving to maturity is w_{11}. Similarly, the genotypes A_1A_2 and A_2A_2 have survival probabilities w_{12} and w_{22}. When allele A_1 has frequency p in the adult population a new born individual in the next generation will have genotype A_1A_1 with probability p^2. So in the next adult generation the frequency of the genotype A_1A_1 will be proportional to p^2w_{11}. (It is not equal, but only proportional to that number because population size is assumed to be fixed.) This gives the following table

Genotype	A_1A_1	A_1A_2	A_2A_2
Frequency in newborns	p^2	$2pq$	q^2
Viability	w_{11}	w_{12}	w_{22}
Frequency of adults	$p^2w_{11}/\bar{w}$	$2pqw_{12}/\bar{w}$	$q^2w_{22}/\bar{w}$

where we have set $q := 1 - p$ and

$$\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22} \tag{7.1}$$

which indicates the mean fitness in the whole population. Only by this constant of proportionality we can assume that population size is constant, i.e.

$$\text{frequency of } A_1A_1 + \text{frequency of } A_1A_2 + \text{frequency of } A_2A_2 = 1.$$

The coefficients w_{11}, w_{12} and w_{22} are called fitness coefficients. It is clear from the above formulas that multiplying all $w_\square$'s with a constant factor does not change genotype frequencies in the following generations.

Another way of writing the above table is by introducing a selection coefficient s and a dominance coefficient h. By this we mean to write

Genotype	A_1A_1	A_1A_2	A_2A_2
Viability	1	$1 - sh$	$1 - s$

which (in the case $s > 0$) means selection against the A_2 allele. Here the effect of selection is more clear: being homozygous for A_2 decreases the viability of an individual by s. Being in an intermediate genotype A_1A_2 means that fitness is not changed to $1 - s$ as for homozygotes but to $1 - sh$. The coefficient h is called the *dominance coefficient* and ranges from $-\infty$ to $1/s$, because viabilities have to be positive numbers. When $h = 0$ then genotypes A_1A_1 and A_1A_2 have the same fitness which means that A_1 is a completely dominant allele. When $h = 1$ then the fitnesses of A_1A_2 and A_2A_2 are the same and so A_2 is dominant. The cases $h < 0$ and $h > 1$ are also of interest. We will come to this later.

Using these notations and conventions there are two different ways to model selection; either by stochastic models (which we will do next) or by deterministic models (which comes afterwards).

Selection in the Wright-Fisher model

The standard Wright-Fisher model was very simple. Every individual had the same chance to be ancestor of any other individual in the next generation. But how can we extend this model to deal with selection? The above tables of fitness coefficients already gives us a hint.

Consider a population with allele frequencies p and $q = 1 - p$ for alleles A_1 and A_2, respectively. When these adults produce zygotes they will be in Hardy-Weinberg equilibrium. However, assume the number of zygotes is very large, much larger than the population size which is N for diploids or $2N$ for haploids. The viability of the zygotes is determined by their genotype and chance. When the offspring grows to adulthood the probability that a randomly chosen adult has genotype A_1A_1 is $p^2 w_{11}/\bar{w}$ with $\bar{w}$ as in (7.1). (Again dividing by $\bar{w}$ is done because of normalization.) For the other genotypes the probabilities are derived analogously.

The Wright-Fisher model only gives allele frequencies but no genotype frequencies. So we have to calculate the allele frequencies from the genotype frequencies. Given the genotype frequencies from (7.1), what is the probability that a randomly chosen allele is of type A_1? This is the probability of either choosing an individual with genotype A_1A_1 or choosing one with A_1A_2 and picking from this individual its A_1 allele which is done with probability $\frac{1}{2}$. When we fill in the appropriate selection and dominance coefficients this gives

$$\tilde{p} = \frac{p^2 + (1 - sh)p(1 - p)}{\bar{w}} = \frac{p(1 - sh) + shp^2}{\bar{w}} \tag{7.2}$$

for the probability that an allele from the next generation picks an A_1 as ancestor and

$$\bar{w} = p^2 + 2(1 - sh)p(1 - p) + (1 - s)(1 - p)^2. \tag{7.3}$$

As individuals are produced independently of each other and given the total number of A_1 alleles in the previous generation t was i this gives the transition probabilities

$$\mathbf{P}[X_{t+1} = j | X_t = i] = \binom{2N}{j} \tilde{p}_i^j (1 - \tilde{p}_i)^{2N-j} = \mathrm{binom}(2N, \tilde{p}_i; j) \tag{7.4}$$

with $\tilde{p}_i$ from (7.2), $p_i = \frac{i}{2N}$, and $\bar{w}$ from (7.3). This formula only holds approximately because it is not true that the haploids are built independently of the diploids. When the first allele of an individual is picked it has probability $\tilde{p}_i$ of being an A_1. But when the second allele is picked the probabilities are already influenced by the first allele. We will discard this dependency and hope that the approximation is good.

The transition probability is a simple binomial distribution, just like in the neutral Fisher-Wright model. We already know from Maths 1.5 the expectation and variance of a binomial distribution and so it is now possible to calculate how selection affects the path of the selected allele. We set $p_t := \frac{X_t}{2N}$ and calculate

$$\begin{aligned}\mathbf{E}[p_{t+1} - p_t | p_t = p] &= \frac{1}{2N}\mathbf{E}[X_{t+1}|X_t = 2Np] - p \\ &= \frac{p(1-sh) + shp^2}{\bar{w}} - p = \ldots = \frac{sp(1-p)\big(1-h+p(2h-1)\big)}{\bar{w}}.\end{aligned} \tag{7.5}$$

Exercise 7.3. Is the right side of (7.5) really correct? Check this! □

For the variance we will only calculate an approximation. We assume that the selection coefficient s is so small that we can assume that $\frac{s}{2N}$ is a small number. As

$$\tilde{p} = p + \mathcal{O}(s),$$

meaning that $\tilde{p}$ and p are different only by a quantity which is proportional to s (and not so $\sqrt{s}$ or so) we immediately have

$$\mathbf{Var}[p_{t+1}|p_t = p] = \frac{1}{2N}\tilde{p}(1-\tilde{p}) = \frac{p(1-p)}{2N} + \mathcal{O}\Big(\frac{s}{2N}\Big) \approx \frac{p(1-p)}{2N} \tag{7.6}$$

which is surprisingly the same variance as in the neutral Fisher-Wright model. So selection influences the expected frequency path but not the variability of the frequency path around this expected curve.

Exercise 7.4. In the beginning of this section we talked about various forms of selection. Consider a case of frequency dependent selection where fitness of an allele is greatest when it is in low frequency. How would you change the Wright-Fisher model to account for this frequency-dependence? □

The fixation probability

What is the probability that a beneficial allele that appears in a population is not lost again but rises to fixation? We give two derivations for this important population genetic quantity. The first one uses a simple argument by Haldane, the second one makes use of the expectation and variance that we have just calculated. For both derivations, we need a so-called Taylor approximation of a function.

John Burdon Sanderson Haldane, 1892–1964; British geneticist, biometrician, physiologist, and popularizer of science who opened new paths of research in population genetics and evolution.
Haldane, R.A. Fisher, and Sewall Wright, in separate mathematical arguments based on analyses of mutation rates, size, reproduction, and other factors, related Darwinian evolutionary theory and Gregor Mendel's laws of heredity. Haldane also contributed to the theory of enzyme action and to studies in human physiology. He possessed a combination of analytic powers, literary abilities, a wide range of knowledge, and a force of personality that produced numerous discoveries in several scientific fields and proved stimulating to an entire generation of research workers. His studies included investigation of the effects of inbreeding and crossbreeding among guinea pigs, animals that he later used in studying the effects of gene action on coat and eye color, among other inherited characters.
He announced himself a Marxist in the 1930s but later became disillusioned with the official party line and with the rise of the controversial Soviet biologist Trofim D. Lysenko. In 1957 Haldane moved to India, where he took citizenship and headed the government Genetics and Biometry Laboratory in Orissa (from Encyclopedia Britannica, 2004).

Maths 7.1. *Let f be a function for which derivatives can be calculated (which applies to almost all functions you know). Under some conditions which are usually met for any point x_0 we can write f as*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n = f(x_0) + f'(x_0)(x - x_0) + \tfrac{1}{2}f''(x_0)(x - x_0)^2 + \dots$$

where we have set $f^{(0)}(x) = f(x)$. The series on the right side is called the Taylor series of f around x_0.

Often Taylor series are used for approximation. Then the function f is approximated by its Taylor series e.g. up to second order; i.e. one discards terms in $(x - x_0)^n$ for $n \geq 3$.

Exercise 7.5. Calculate the Taylor series of the function $f(x) = e^x$ around 0 up to the order 1. Compare your findings with Maths 1.1. Do you already see that

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}?$$

□

Haldane's approximation Let π_{fix} be the fixation probability of a single copy of a beneficial allele with selection coefficient s and dominance coefficient h. As long as the mutant is rare, it will only appear in heterozygotes, which have a selection coefficient hs. We assume a Wright-Fisher population with an (approximately) Poisson distributed offspring number. Beneficial heterozygotes have on average $1 + hs$ offspring.

We know that in some distant future, the allele will either fix or get lost. The probability of fixation is therefore equivalent to the probability of not getting lost. Obviously, the probability of not getting lost for a specific allele is the probability that it has offspring at all *and* at least one of the offspring-alleles is not lost from the population (but has descendents in the distant future). As long as the beneficial allele is rare, the fixation probability (probability of not getting lost) for each copy of the allele in the offspring generation is the same as the original fixation probability. We therefore obtain the following recursion

$$\begin{aligned}\pi_{fix} &= \sum_k \frac{\exp[-(1+hs)](1+hs)^k}{k!}\Big(1-(1-\pi_{fix})^k\Big) \\ &= 1-\exp[-(1+hs)]\cdot\exp[(1+hs)(1-\pi_{fix})] = 1-\exp[-\pi_{fix}(1+hs)]\end{aligned}$$

Under the assumption of a small fixation probability, we can approximate the exponential by the first terms of its Taylor expansion $\exp[-\pi_{fix}(1+hs)] \approx 1-\pi_{fix}(1+hs)+(\pi_{fix}(1+hs))^2/2$ and obtain

$$\pi_{fix} \approx \frac{2hs}{(1+hs)^2} \approx 2hs. \tag{7.7}$$

Since a value of hs of a few percent is already considered strong selection, we see that the large majority of all new beneficial alleles will get lost from the population by drift.

Derivation of Kimura's formula We can also derive a slightly more precise formula that also works if the initial frequency of the beneficial allele is not very low. Assume that the allele is already present at frequency p. If the allele were neutral the probability that it eventually fixed is p (see Section 4.3). But what if the allele is beneficial? For simplicity we set $h = \frac{1}{2}$. Then selection is additive because the fitness is determined by the number of selected alleles an individual carries. Define

$$\pi(p) = \mathbf{P}[\text{fixation}|p_0 = p],$$

i.e. $\pi(p)$ is the fixation probability of the allele given it is in frequency p today.

The basic equation we use is that

$$\pi(p) = \sum_{\Delta p} \mathbf{P}[p_1 - p = \Delta p]\pi(p+\Delta p) = \mathbf{E}[\pi(p+\Delta p)|p_0 = p]. \tag{7.8}$$

The second equality is just the definition of an expectation. The first equality means that the fixation probability given the frequency p is transported to the next generation. When in the next generation the allele has frequency $p+\Delta p$, which happens with probability $\mathbf{P}[p_1 - p = \Delta p]$ the new fixation probability is $\pi'(p+\Delta p)$. But necessarily $\pi = \pi'$ because the fixation probability does not depend on the number of the generation.

For $\pi(p+\Delta p)$ we use a Taylor series approximation around p up to order 2. We have

$$\pi(p+\Delta p) = \pi(p) + \pi'(p)\Delta p + \tfrac{1}{2}\pi''(p)(\Delta p)^2. \tag{7.9}$$

Combining (7.8) and (7.9) we have approximately

$$\pi(p) = \pi(p) + \pi'(p)\mathbf{E}[\Delta p] + \tfrac{1}{2}\pi''(p)\mathbf{Var}[\Delta p]$$

and so

$$0 = \tfrac{1}{2}sp(1-p)\pi'(p) + \tfrac{1}{2}\frac{p(1-p)}{2N}\pi''(p)$$

or

$$\pi''(p) = -2Ns\pi'(p).$$

This is a differential equation for $\pi(.)$. However it is very simple. First, every function f for which we know that $f' = af$ for some number a satisfies $f(x) = ce^{ax}$. So we know taking π' for the function f that

$$\pi'(p) = C\exp(-2Nsp).$$

But then with $C' = C/2Ns$

$$\pi(p) = \int C\exp(-2Nsp)dp = C'\exp(-2Nsp) + D$$

for some number D. The only thing that remains to be calculated are the numbers C' and D. This is done by using boundary conditions. There are two numbers we already know about $\pi(.)$. First $\pi(0) = 0$ as an allele cannot be fixed when it is not present in a population (and, of course, when no mutation is assumed). Second, $\pi(1) = 1$ as in this case the allele is already fixed in the population. This gives two equations, namely

$$C' + D = 0, \qquad 1 = C'\exp(-2Ns) + D.$$

Plugging the first into the second equation gives

$$-1 = C'\big(1 - \exp(-2Ns)\big), \qquad C' = -\frac{1}{1-\exp(-2Ns)}, \qquad D = \frac{1}{1-\exp(-2Ns)}$$

and so

$$\pi(p) = \frac{1-\exp(-2Nsp)}{1-\exp(-2Ns)}. \tag{7.10}$$

Exercise 7.6. Assume a new mutant enters the population which has a fitness advantage of s. Initially its frequency is $1/2N$. Can you re-derive Haldane's approximation (7.7) from (7.10)? □

Exercise 7.7. We made several approximations in our calculation of the fixation probability (which ones?). It is certainly worth checking (7.10) numerically. In addition to that we can also check (7.7) and see when it breaks down.

For the simulations we use `wf.freq()` from the `R`-package. Here the above described Wright-Fisher model with selection and fitnesses 1, $1-sh$ and $1-s$ is implemented. You will use the parameter `batch` to study the fixation probability of a newly arisen beneficial allele. To do this always set `init.A` to `1/2N`.

1. Use `N`=500. Simulate 1000 runs and plot the average curve using

```
>ret<-wf.freq(N=500,init.A=0.001,s=0.01,stoptime=2000,batch=1000)
>plot(ret, what="freq.A")
```

   As eventually an allele is either lost or fixed the Y-axis shows you the frequency of runs where the beneficial allele fixed. To see the exact value in how many of your runs the allele A fixed, use

```
>mean(ret$freq.A[2000,])
```

   Use $s = 0.2, 0.05, 0.01, 0.001$ and simulate the fixation probability. Compare your values with (7.10) and (7.7). Where do you see the greatest deviance? How much do you trust in your runs for small values of s?

2. Make a plot out of the values. Compare your plots with other members of the class. You should get a hint in which parameter range to use (7.10) and (7.7).

□

Deterministic models

As we saw in the Wright-Fisher model the expectation and the variance of the change in allele frequency were given by (7.5) and (7.6). We saw that the expectation was of the order $\mathcal{O}(s)$ whereas the variance was of the order $\mathcal{O}(1/2N)$. When selection is strong or population size is large, i.e. when $s \gg 1/2N$ or $2Ns \gg 1$, we can ignore the variance, and thus forget about the stochasticity, i.e. forget about genetic drift and use a deterministic model instead to describe the changes in allele frequencies where

$$p_{t+1} - p_t = sp_t(1-p_t)\frac{1-h+p_t(2h-1)}{\bar{w}}.$$

Another approximation is to use continuous time. That means we use a differential equation, i.e.

$$\frac{dp}{dt} = sp(1-p)\frac{1-h+p(2h-1)}{\bar{w}}. \tag{7.11}$$

Exercise 7.8. Again it was Sewall Wright who claimed (already in the thirties) that

$$\frac{dp}{dt} = \frac{p(1-p)}{2\bar{w}} \frac{d\bar{w}}{dp}.$$

Can you check if he was right?

What can you read from this equation? E.g. what is a sufficient and necessary condition for the allele frequency to increase in terms of the average fitness? Exactly when does the frequency not change any more?

□

7.3 Examples

There are some main examples we will deal with. Depending on the selection and dominance coefficients the solution of the differential equation (7.11) has different properties. The function `selectionDet()` is an implementation of the differential equation (7.11). We will be studying several cases.

$h = 0$	selection against a recessive allele
$h = 1$	*a* selection against a dominant allele
$0 < h < 1$	incomplete dominance
$h = \frac{1}{2}$	additive selection
$h < 0$	overdominance
$h > 1$	underdominance

When $h = 0$, which means that genotype A_1A_1 is as fit as genotype A_1A_2 meaning that the disadvantageous A_2-allele is recessive. When $h = 1$ the roles of A_1 and A_2 are interchanged and so A_2 must be dominant and A_1 recessive. Any point between these two extreme cases, i.e. $0 < h < 1$ is denoted by incomplete dominance. Here the fitness of the heterozygote is somewhere between the homozygotes. But also the case of a fitter heterozygote must be assumed. Here $h < 0$ which means that the heterozygote has fitness greater than 1. The last case is when the heterozygote is less fit than the two homozygotes which occurs for $h > 1$.

We will be dealing with all cases using biological examples using our `R`-package. To make e.g. a plot of the frequency curve of an allele with $s = 0.01, h = 0.5$, initial frequency 0.1 for the first 500 generations, you may type in one command

```
>plot(selectionDet(init.A=0.1,s=0.01,h=0.5,stoptime=500),what="freq.A")
```

Exercise 7.9. The allele for the human disease phenylketonuria (PKU) is recessive and causes the body to be unable to process phenylalanine in the homozygous state. Consider a deterministic model in which the PKU allele is initially rare ($p_0 = 0.05$) and in which the PKU phenotype has a fitness of 0.25 relative to the normal phenotype.

1. Which case of the above table applies in this case?

2. What happens to the PKU allele frequency over time? How many generations would it take for the PKU allele to reach a frequency of 0.01? (By adding `plot(..., xlim=c(90,110), ylim=c(0.98,1))` or something more appropriate you see a zoomed version of your original plot. You might also want to use `abline(h=0.99)` to see when the PKU-allele has reached a frequency of 0.01.)

3. Now consider what would happen if better screening techniques were developed such that more individuals with PKU could receive treatment immediately, and the fitness of the PKU phenotype increased to 0.80 relative to the normal phenotype which has fitness 1. The initial frequency of PKU allele is still 0.05. In this situation, how many generations would it take for the PKU allele to reach frequency 0.01?

□

Exercise 7.10. Huntington's disease is inherited as a Mendelian dominant phenotype. The disease is characterized by neural degeneration that often does not set in until after the individual has passed child-bearing age.

1. Which case of the table on page 104 applies in this case?

2. Consider a deterministic model in which the Huntington's allele initially occurs in low frequency ($p_0 = 0.05$) and has a selection coefficient of only 0.20 due to the late onset of the disease. How many generations would it take for the Huntington's allele to reach frequency 0.01?

□

Exercise 7.11. 1. Compare the length of time required for the deleterious allele to reach frequency 0.01 in the case of the PKU allele (Exercise 7.9) and in the case of the Huntington's allele (Exercise 7.10). In which case does it take longer? Explain why. The situations are the same except for the dominance of the allele.

HINT: Think about what selection *sees* in each case and about the frequency of homozygotes of rare alleles.

□

Exercise 7.12. 1. Consider a situation in which a favored, dominant allele A_1 is initially rare in a population with $p_0 = 0.05$. Explore some different values for the selection coefficient against the recessive homozygote until you get a feeling for how the allele frequency changes over time. Choose a value for s that is between 0.5 and 0.2 and set `stoptime` to 150 generations. Sketch the graph of how the frequency of the dominant allele changes over time.

2. Now consider the same situation as above but with a favored recessive allele A_1 (i.e. use the same selection coefficient and initial p, and just change the fitness of the heterozygote).

   Sketch the dominant case from Exercise 7.10 and the recessive case from Exercise 7.9.

3. Initially (when the favored allele A_1 is still rare), in which case (A_1 dominant or A_1 recessive) is selection more efficient at increasing the frequency of the favored allele?

4. Towards the end of the process (when the deleterious A_2 allele is rare), in which case is selection more efficient at increasing the frequency of the favored allele (i.e. decreasing the frequency of the deleterious allele)?

5. Explain why these two selective regimes have different effects on allele frequencies over time (i.e. why the curves are shaped differently).

□

In the two cases, A_1 dominant and A_1 recessive when the A_1 allele manages to prevail in the population it will eventually fix. This is also true for the case of incomplete dominance. However this situations differs in the case of overdominance. Here the heterozygote is fitter than both homozygotes. Therefore the population will not eliminate the A_2 allele because it will still be present in most fit heterozygotes. But at which frequency will it be?

This can e.g. be calculated from (7.11). The allele frequencies do not change any more if

$$1 - h + p_*(2h - 1) = 0 \qquad \text{or} \qquad p_* = \frac{1-h}{1-2h}. \tag{7.12}$$

Here p_* is called the equilibrium of allele frequencies. Clearly p_* is between 0 and 1 if either $h < 0$ or $h > 1$, i.e. in the cases of over- and underdominance.

Exercise 7.13. Sickle cell anemia has been described as an example of heterozygote advantage (where the heterozygote has a higher fitness than either homozygote). Explore different parameter values that represent this situation.

1. What happens to allele frequencies over time for different values of s? What is the equilibrium allele frequency?

2. Does changing the initial allele frequency affect the value of the allele frequencies at equilibrium? If so, how?

3. What does it mean to have reached an equilibrium?

4. Now examine the average fitness of a population $\bar{w}$ undergoing selection at a locus with heterozygote advantage. (To see $\bar{w}$, add `plot(..., what="w.bar")` to your command. You might also want to add `plot(..., ylim=c(0,2))` to expand the

y-axis.) Experiment with different values for s and h, keeping $h < 0$. In the case of heterozygote advantage, does the population ever reach its maximal fitness of $1-hs$? Briefly explain why selection does not increase the average fitness of the population to its maximum in this case.

5. Check (numerically or using a calculation) that the average fitness in equilibrium is

$$\bar{w}^* = 1 + s\frac{h^2}{1-2h}.$$

□

Also in the case of underdominance, i.e. $h > 1$ meaning that the heterozygote has a lower fitness that both homozygotes (7.12) gives an equilibrium value of p between 0 and 1.

Exercise 7.14. Now consider a case of underdominant selection.

1. Explore different parameter values for a while.

2. Now vary the initial allele frequency, keeping the fitness the same. Does the value of this parameter affect the outcome of the model (i.e. does it affect which allele eventually reaches a frequency of 1 or whether either allele reaches that frequency)? If so, explain how, and describe the possible outcomes of the model.

3. Again (7.12) gives an equilibrium in this case. Set the initial frequency exactly to this value. What happens if you change this frequency a little bit? Can you explain this behavior? This behavior is described as an instable equilibrium. Can you explain why?

4. Plot also $\bar{w}$. As you see it always increases. However its maximal value is not always 1. What happens in these cases?

□

8 Selection and polymorphism

Now we are going to look at how selection can influence variation and how variation in turn can influence selection. The first subsection deals with negative selection, it shows that deleterious mutations can be kept at an equilibrium frequency in a population and how these deleterious mutations effect the mean fitness of the population. The second subsection introduces the idea that, at least in simple models, mean fitness always increases. The third subsection is again about deleterious mutations and the distribution of the number of mutations that individuals carry. It also shows how the minimum number of deleterious mutations carried by an individual in a population may increase over time due to an effect that is named *Muller's Ratchet*. Finally the fourth subsection is about the effect of strong positive selection on linked neutral variation, this effect is called hitchhiking.

8.1 Mutation-Selection balance

Let us consider deleterious mutations, or, in other words, alleles that have a negative fitness effect. Individuals carrying such mutations or alleles will have, on average, fewer offspring than the other individuals in the population. Through this negative selection the mutations will eventually disappear from the population if mutation does not create new alleles of this type. If mutation does create the allele with lower fitness at rate u, that is

$$A \longrightarrow a \text{ at rate } u$$

then a dynamic equilibrium can be reached. Mutations enter the population at a certain rate, depending on mutation, and will be purged from the population due to selection. The equilibrium is called mutation-selection balance. Or if we include stochasticity, mutation-selection-drift balance.

We assume only viability selection. This means selection acts between the zygote and adulthood. Mutations, however, happen at the production of gametes, before building the zygotes. Now if the frequency of the wildtype allele A is p in the adult population, then in their zygotes the frequency of allele A will be

$$p' := p(1-u).$$

.

Remember, from these zygotes, we calculated the effect of selection on the next generation in (7.4) in the last section. We can do the same calculation in our present context. We have (from the Wright-Fisher model)

$$\mathbf{P}[X_{t+1} = j | X_t = 2Np] = \text{binom}(2N, \tilde{p}'; j)$$

where $\tilde{p}'$ is - according to (7.2) - the probability that an individual in the next generation will choose a parent with genotype A, given the set it can choose from has a frequency of p' of the A alleles. This probability is the normalized fitness of the A individuals times the

frequency of the A allele in the zygotes. To normalize the fitness we need the mean fitness $\bar{w}$.

$$\bar{w} = (p')^2 + 2(1-sh)p'(1-p') + (1-s)(1-p')^2 = 1 - 2shp'(1-p') - s(1-p')^2. \quad (8.1)$$

Approximately we can say that by mutation the allele frequency is decreased from p to $(1-u)p$, so the difference is $-up$. So if an equilibrium is to be reached the increase should be up. We have already calculated the increase in frequency as

$$\frac{sp(1-p)(1-h+p(2h-1))}{\bar{w}}$$

In equilibrium we must therefore have

$$up = \frac{sp(1-p)(1-h+p(2h-1))}{\bar{w}}.$$

Discarding terms of order $u(1-p)^2$ and $us(1-p)$ we approximately have that

$$u = sh(1-p)(2p-1) + s(1-p)^2. \quad (8.2)$$

Also ignoring terms $s(1-p)^2$, assuming that $p \approx 1$ such that $2p-1 \approx 1$ we find that

$$1 - p \approx \frac{u}{sh}. \quad (8.3)$$

Exercise 8.1. In the case of a recessive disadvantageous allele we have $h = 0$ and so the above equation breaks down. Can you calculate an approximation of the mutation-selection balance in this case? □

In equilibrium we can also ask what the average fitness of the population will be. Plugging (8.3) in (8.1) we find, ignoring differences between p and p' on $\bar{w}$, that

$$\bar{w} = 1 - 2sh\Big(1 - \frac{u}{sh}\Big)\frac{u}{sh} - s\frac{u^2}{(sh)^2} \approx 1 - 2u$$

where we have ignored some terms.

The difference of the maximum fitness to the average fitness in the population is called the *genetic load*. The genetic load is defined as

$$L = w_{\max} - \bar{w}$$

and so in our case is

$$L = 2u, \quad (8.4)$$

which is, surprisingly, independent of s and h! The result that the genetic load is independent of s and h but only depends on the mutation rate is called the *Haldane-Muller*

Hermann J. Muller, 1890–1967, American geneticist, is best remembered for his demonstration that mutations and hereditary changes can be caused by X rays striking the genes and chromosomes of living cells. His discovery of artificially induced mutations in genes had far-reaching consequences, and he was awarded the Nobel Prize for Physiology or Medicine in 1946.
Muller attended Columbia University where his interest in genetics was fired first by E.B. Wilson, the founder of the cellular approach to heredity, and later by T.H. Morgan, who had just introduced the fruit fly Drosophila as a tool in experimental genetics. The possibility of consciously guiding the evolution of man was the initial motive in Muller's scientific work and social attitudes. His early experience at Columbia convinced him that the first necessary prerequisite was a better understanding of the processes of heredity and variation.
He produced a series of papers, now classic, on the mechanism of crossing-over of genes, obtaining his Ph.D. in 1916. His dissertation established the principle of the linear linkage of genes in heredity. The work of the Drosophila group, headed by Morgan, was summarized in 1915 in the book *The Mechanism of Mendelian Heredity*. This book is a cornerstone of classical genetics.
Muller was a socialist, and he initially viewed the Soviet Union as a progressive, experimental society that could pursue important research in genetics and eugenics. But by this time the false doctrines of the biologist T.D. Lysenko were becoming politically powerful, bringing to an end valid Soviet scientific research in genetics. Muller fought Lysenkoism whenever possible, but he ultimately had to leave the Soviet Union. Then he worked in Edinburgh and several universities in the USA (adapted from Encyclopedia Britannica, 2004).

principle. This principle says that the mean fitness of the population doesn't depend on the effect of single mutations, but only on the rate at which they occur in the population. To understand this, think of a population where every mutation has a large fitness effect. In such a population an individual carrying a deleterious mutation would most likely not leave any offspring and the mutation would be kept at a very low frequency. The effect on the individual is large, but the effect on the population is not so large. On the other hand, if you would consider a population where mutations would have very small effects, then individuals with these mutations would most likely leave as much offspring as the other individuals and therefore the frequency of the mutation can be higher in the population. The effect on the individual would be small, but the effect on the population would be the same. The reason that the genetic load is independent of h is the same as for s. Since deleterious mutations are expected to be rare and only occur in heterozygotes, s and h always occur together in the equations anyway, as a compound parameter.

Exercise 8.2. Let us check the Haldane-Muller principle numerically. We use the function `wf.freq()` of the `R`-package. Here you can also set mutation parameters `uA2a` which is the rate for an A to mutate to an a and `ua2A` for the other direction. Set $N = 500$.

1. Take $u = 0.01$. Vary s from 0.01 to 0.5 and h from 0 to 2 to see the parameter range where the Haldane-Muller principle is valid. Use `batch=100` to obtain the average

of 100 runs of the simulation. Make a plot that compares actual values with the theoretical prediction.

2. In 1. you will find some parameter combinations where the Haldane-Muller principle it is not applicable. Can you explain what happens in these ranges?

3. For $h = 0$ you calculated in Exercise 8.1 the equilibrium frequency for a dominant beneficial allele. Is the genetic load also independent of s and h in this case? Compare your calculation with the values you obtained in 1.

□

8.2 The fundamental Theorem of Selection

Ronald Fisher stated that 'the rate of increase in fitness of any organism at any time is proportional to its genetic variance in fitness at that time.' This has since been called the *Fundamental Theorem of Selection*. Although the term *genetic variance* comes from quantitative genetics with which we are not dealing in this course, we can at least conclude that the rate of increase in fitness must be positive because the variance in fitness has to be positive. So in other words, fitness will always go up. A useful way to see this process is to imagine the fitness landscape as a hilly landscape, where a population is always moving towards the top of a hill.

In Exercise 7.8 we found an equation relating the mean fitness of an allele to its evolution in time. Let's assume there is a stable equilibrium point p_* for p with $0 \leq p_* \leq 1$, so either partial dominance or overdominance. As long as $p < p_*$ we have $\frac{dp}{dt} > 0$ and thus $\frac{d\bar{w}}{dt} > 0$ so fitness will increase. When $p > p_*$ then $\frac{dp}{dt} < 0$ and so $\frac{d\bar{w}}{dp} < 0$ leading to

$$\frac{d\bar{w}}{dt} = \frac{d\bar{w}}{dp}\frac{dp}{dt} > 0.$$

So we see that the average fitness always increases in this model.

Exercise 8.3. What will happen if selection is underdominant, i.e. $h > 1$? □

Exercise 8.4. Draw $\frac{d\bar{w}}{dt}$ against $\bar{w}$ in the case of partial or overdominance. To do this, proceed (e.g. for $h = 0.9$) as follows:

```
>ret<-selectionDet(init.A=0.1, s=0.1, h=0.9, stoptime=150)
>plot(ret$w.bar[1:149], ret$w.bar[2:150]-ret$w.bar[1:149],xlab="", ylab="")
>title(xlab=expression(paste(bar(w))))
>title(ylab=expression(paste(d, bar(w), "/dt")))
```

The qualitative evolution of $\bar{w}$ can be read from this diagram which is often referred to as a *phase diagram*. Draw the direction of the movement of $\bar{w}$ on the x-axis in the diagram.

Unfortunately the fundamental theorem of selection is only true for a very simple model with constant selection. The Fundamental Theorem of Selection was under heavy debate and today is assumed not to be generally true (and thus not a theorem). Let us make an example. We have already seen that the theorem holds for the simple model, so in order to show that it is not general we need to look at a more involved one.Let us consider two partially linked loci A and B that both contribute to the fitness of an individual. Both loci have two alleles A/a and B/b. Furthermore assume that the fitness of AB is higher than that of any other haplotype and the population starts off with a large fraction of AB's. Then by recombination these genotypes are broken up leading to less fit individuals. Consequently when the initial frequency was high enough the average fitness will decrease and so the Fundamental Theorem does not hold. Other examples involve changing fitness landscapes. Suppose that a population needs some resource, and as the population gets better at using the resource, the resource gets depleted, what was a peak in the landscape becomes a valley as soon as the population arrives there. Mean fitness may go down in this case.

Exercise 8.5. Can you think of yet another example? □

8.3 Muller's Ratchet

Deleterious mutations that accumulate in a population do not only lower the mean fitness of a population if they occur at an equilibrium frequency of mutation- selection and drift. Another question is how many individuals in a population carry zero, one or two etc. deleterious mutations. Of course deleterious mutations can be removed from the population by selection. Under some circumstances, i.e. when drift is strong enough, it can happen that the class of individuals with no deleterious mutations goes extinct. This is described as one *click* of *Muller's Ratchet*. If this happens repeatedly it can be dangerous for the population because it can eventually drive it extinct. If Muller's Ratchet clicks, the maximal fitness is a little bit less than before. It is assumed that in humans every newborn carries three or four newly arisen harmful mutations. The reason why they do not cause a major problem for the human population are twofold. First, they are often almost recessive and so they don't have a large fitness effect. Second, sex helps: by recombination mutations are reordered and thus it is possible that individuals have fewer mutations than both of their parents. This is in contrast to asexual populations, where recombination is not possible and as soon as the ratchet has clicked there is no way to recreate the class of zero mutations.

We will approximate the frequency distribution of the number of deleterious mutations that individuals carry. We assume the following: each mutation has a multiplicative fitness effect of $(1-s)$ so if an individual carries i mutations its fitness is $(1-s)^i$. During reproduction new mutations accumulate. In our model we assume that the number of new mutations each generation in an individual is a Poisson number with parameter u.

Exercise 8.6. Let us see how fast mutations are accumulated depending on the parameters N, s and u.

1. Use the function `seqEvoMullersRatchet()` to see how mutations are accumulated. Running e.g. `seqEvoMullersRatchet(wait=-1)` lets you see how the sequence patterns change in each generation. For `wait=0.2` you see a new picture every 0.2 seconds.

2. Fix $N = 50$ and use the `plot=FALSE` option. How often does the ratchet click for $u = 1, 0.5, 0.3, 0.1$ and $s = 0.5, 0.3, 0.1, 0.05$ in 100 generations?

   To efficiently simulate this you might want to do the following:

```
>u<-c(1,0.5,0.3,0.1)
>s<-c(0.5,0.3,0.1,0.05)
>rate<-matrix(0, ncol=4, nrow=4)
>for(i in 1:4)
+for(j in 1:4)
+rate[i,j]<-seqEvoMullersRatchet(N=50, u=u[i], s=s[j],
+stoptime=100, plot=FALSE)
```

   Why does the number of clicks increase with increasing u and decreasing s?

3. Haigh (1978) estimated that the average time between clicks of the ratchet is approximately $N \cdot e^{-u/2}$. How well do your results reflect his approximation?

□

In the last exercise you saw that the rate of the ratchet was not well approximated. Today it is still an open problem to determine the rate at which deleterious mutations accumulate, depending on the model parameters N, u and s.

The ratchet can operate best if genetic drift operates; hence, the ratchet cannot work in large populations. In the limit of an infinite population, we next compute the distribution of the number of deleterious alleles in the population after a long time. So we would like to know the number of individuals that has 0, 1, 2, etc deleterious mutations. To make this a feasible task we rely on the equilibrium that will be reached to be stable. When we can guarantee that this stable equilibrium exists, all we have to do is to find the equilibrium and show that it does not change any more. We will *try* a Poisson distribution for the number of deleterious alleles that is carried by the individuals in the population with some (yet unspecified) parameter λ.

If we assume that at some time t this probability that one individual carries i mutations is

$$p_i(t) = e^{-\lambda} \frac{\lambda^i}{i!}.$$

As mutations accumulate during reproduction we will have to compute the distribution of a sum of Poisson random variables.

Maths 8.1. *Let $X \sim pois(\lambda)$ and $Y \sim pois(\mu)$ two independent random variables. Then*

$$\begin{aligned}\mathbf{P}[X+Y=k] &= \sum_{i=0}^{k} \mathbf{P}[X=i, Y=k-i] = \sum_{i=0}^{k} e^{-\lambda} e^{-\mu} \frac{\lambda^i}{i!} \frac{\mu^{k-i}}{(k-i)!} \\ &= \frac{1}{k!} e^{-(\lambda+\mu)} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \mu^{k-i} = \frac{1}{k!} e^{-(\lambda+\mu)} (\lambda+\mu)^k,\end{aligned}$$

so $X+Y \sim pois(\lambda+\mu)$.

After accumulation of new mutations the number of deleterious mutations is therefore Poisson distributed with parameter

$$\lambda' = \lambda + u,$$

so after accumulation of new mutations (i.e. when forming gametes) in the next generation

$$p_i'(t+1) = e^{-\lambda'} \frac{(\lambda')^i}{i!}.$$

Then reproduction uses selection to increase the frequency of genotypes with only a few mutations. Writing p_i for $p_i(t)$ and p_i' for $p_i'(t+1)$, we have

$$\begin{aligned}p_i(t+1) &= \frac{p_i'(1-s)^i}{\sum_{j=0}^{\infty} p_j'(1-s)^j} = \frac{e^{-(\lambda+u)}(\lambda+u)^i(1-s)^i}{i! e^{-(\lambda+u)} \sum_{j=0}^{\infty} \frac{(\lambda+u)^j(1-s)^j}{j!}} \\ &= e^{-(\lambda+u)(1-s)} \frac{(\lambda+u)^i(1-s)^i}{i!} = \mathrm{pois}((\lambda+u)(1-s))(i).\end{aligned}$$

Here we have used the formula for e^x from Exercise 7.5. So under the above assumptions $p_i(t+1)$ is again Poisson distributed. In equilibrium the parameters of the Poisson distributions for $p_i(t)$ and $p_i(t+1)$ must coincide. Thus

$$\lambda = (\lambda+u)(1-s) \qquad \text{or} \qquad 0 \approx u - \lambda s$$

which gives

$$\lambda^* = \frac{u}{s}.$$

So in equilibrium the distribution of the number of deleterious alleles a randomly picked individual carries is Poisson distributed with parameter λ^*. This is the same as to say that the frequency of individuals which have i mutations is

$$e^{u/s} \frac{(u/s)^i}{i!}.$$

So for $i=0$ the frequency is $e^{-u/s}$ and the absolute number is $Ne^{-u/s}$.

Exercise 8.7. In this exercise you will use `mullersRatchet()` from the `R`-package to see how the equilibrium distribution of deleterious alleles is approached.

1. The frequencies for the first 20 classes and first 50 generations for $u = 0.1$ and $s = 0.05$ are generated using

   ```
   >ret<-mullersRatchet(u=0.1,s=0.05,init.freq=1,stoptime=50,n.class=20)
   ```

   Here, the starting position was a population without any deleterious mutation and thus $p_0 = 1$. The frequency of class $j-1$ in generation i can now be accessed using `ret$freq[i,j]`.

2. Typing `ret$freq[2,1]` tells you that the frequency of the best class in generation 2 is approximately 90%. Can you compute using the above formulas that this value is correct? What is the equilibrium frequency of the class carrying no deleterious mutation?

3. To see what the frequencies after 50 generations look like compared to the equilibrium, we use `plot(ret)`. Are the frequencies already in equilibrium? Is there a big difference after 500 generations?

4. To see how the equilirbium is approached, we use

   ```
   >for(i in 1:50) plot(ret, i)
   ```

   Use different starting distributions to see if the Poisson equilibrium is always attained.

5. Explain in your own words what an equilibrium distribution is and why it means that the Poisson distribution with parameter $\frac{u}{s}$ is such a thing in the present context.

6. We said that the Poisson equilibrium must be valid if the population size is large such that genetic drift cannot operate. Assume $N = 10^6$, $u = 1$ and $s = 0.05$. How many *individuals* would you expect carrying no mutations in equilibrium, i.e., under the Poisson distribution? Would you expect that such an N is already large enough such that the ratchet does not click?

□

As you have seen, the frequency of the zero-mutations-class can become low under certain parameter values. If you would have a finite population size, this low frequency means a small number of individuals. And because in reality offspring number is a random variable, the few individuals may fail to reproduce and the zero-mutations-class may die out. In a population without recombination this can not be reversed (unless by a back mutation, which is expected to be very rare). If this happens the ratchet has clicked and the 'equilibrium' distribution will have shifted one mutation.

Exercise 8.8. Muller's ratchet deals with the decrease in fitness. In (8.4) we computed the genetic load $L = 2u$, which also describes a decrease in fitness. What are the differences in the assumptions of Muller's ratchet and the calculation with genetic load? Especially, answer the following questions for both models:

1. Is the maximal or mean fitness affected?
2. Does recombination play an important role in the model?
3. Is the ratio u/s or the mutation rate u the most important factor in the analysis of the model?
4. Is the effect predominant in finite populations?

□

8.4 Hitchhiking

Detecting where positive selection acts in a genome is one of the major goals when studying evolution at the molecular level. Positive selection leads to the fixation of mutations that lead to new adaptations (i.e. fitness increase) in a population. Of course we do not only want to know where it acts in the genome but also how. But here we will look at a way to detect the location of positive selection.

In 1974 John Maynard Smith[5] and John Haigh wrote a paper on the effect of positive or directional selection on heterozygosity at a linked neutral locus (Maynard Smith and Haigh, 1974). They considered (roughly) the following situation: A mutation happens that changes b into B. B has a selective advantage over b. Close to the b-locus there is a neutral a-locus. This locus is polymorphic, which means that there is an a allele and an A allele (that could be a SNP, or a microsatellite or an indel). The mutation from b to B occurred on a chromosome carrying an a-allele. If the A/a and the B/b locus are very close together (which means tightly linked), the B mutation will drag along the a-allele to fixation. This will lead to lower polymorphism at the a-locus, an effect that is called a selective sweep. Selection happens at the gametic stage and if we assume that there are no dominance effect, a gamete carrying a B has a relative fitness of $1 + s/2$ and a gamete carrying a b has a relative fitness of 1.

One example of a region of decreased polymorphism was found in Schlenke and Begun (2003). A region in the genome of *Drosophila simulans* was known to carry alleles responsible for *DDT resistance*. So these loci should have been under strong selection in the 50's 60's and 70's of the last century. The pattern that was observed here is given in Figure 8.1

[5] 1920-2004; British evolutionary biologist and geneticist. Originally an aeronautical engineer during the Second World War, he then took a second degree in genetics under the great J.B.S. Haldane. Maynard Smith was instrumental in the application of game theory to evolution and theorised on other problems such as the evolution of sex and signaling theory. (from Wikipedia)

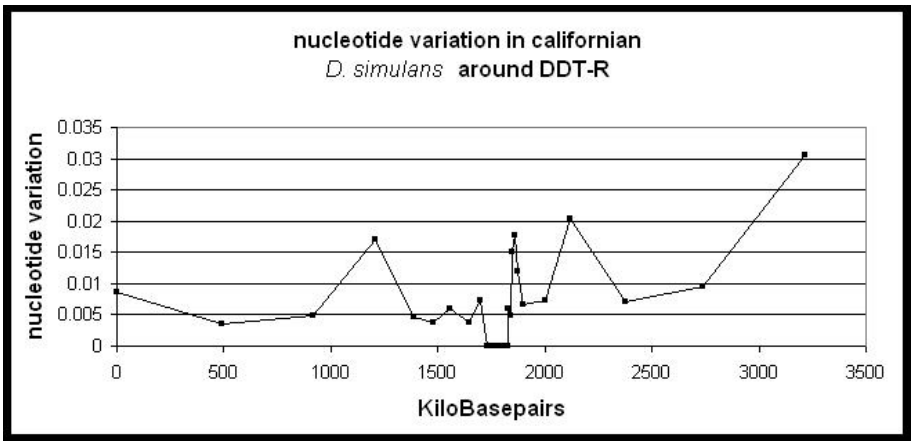


Figure 8.1: A region of reduced polymorphism as found in Schlenke and Begun (2003).

Let's model hitchhiking mathematically. We are using a deterministic model, which should not be a problem if we assume that selection is strong compared to drift. This is justified for large s because we already calculated that genetic drift is much weaker than selection for large selection coefficients (see Section 7). Let p_t be the allele frequency of B at time t and $p_0 = \frac{1}{2N}$ which means that initially one B enters the population. Additionally the frequency of B is given by the differential equation

$$\dot{p} = \tfrac{1}{2}sp(1-p). \tag{8.5}$$

Exercise 8.9. Can you find the differential equation from Section 7 that boils down to (8.5)? □

Exercise 8.10. Can you calculate using Maths 8.2 that

$$p_t = \frac{p_0}{p_0 + (1-p_0)e^{-st/2}} \tag{8.6}$$

solves the differential equation (8.5) with the correct start value? □

Assume you do not know the solution of (8.5) given in the exercise. There is still a way to find it.

Maths 8.2. *Some differential equations can easily be solved. Take e.g.*

$$\frac{dg(x)}{dx} = f(x)g(x).$$

Then formally rewriting this as

$$\frac{dg}{g} = f(x)dx$$

and integrating gives

$$\log g(x) = \int \frac{1}{g} dg = \int f(x)dx + C$$

for some constant C. If the integral on the right side can be solved explicitely you can now solve this equation for g. The constant C is used to adjust some already given value of g.

Exercise 8.11. Can you apply Maths 8.2 to find the solution of (8.5) given in Exercise 8.10? □

A first question concerning hitchhiking could be: how long does it take for the B allele to fix in the population? As our model is deterministic we ignore drift and the solution to (8.5), which was given by (8.6), has 1 only as an asymptote. So we ask for the time τ it takes to go from a frequency p_0 to $1 - p_0$ using $p_0 = \frac{1}{2N}$. We can use the same trick as in Maths 8.2 and rewrite

$$\frac{2dp}{sp(1-p)} = dt$$

and integrate

$$\int_{p_0}^{1-p_0} \frac{2}{sp(1-p)} dp = \int_0^{\tau} 1dt = \tau$$

and so

$$\begin{aligned} \tau &= \frac{2}{s} \int_{p_0}^{1-p_0} \Big(\frac{1}{p} + \frac{1}{1-p}\Big) dp = \frac{2}{s} \big(\log p - \log(1-p) \big) \Big|_{p_0}^{1-p_0} \\ &= \frac{2}{s} \log \Big(\frac{p}{1-p} \Big) \Big|_{p_0}^{1-p_0} = \frac{4}{s} \log \Big(\frac{1-p_0}{p_0} \Big) \approx \tfrac{4}{s} \log(2N). \end{aligned} \tag{8.7}$$

When s is not too small this is much smaller that $4N$ which is the expected time of fixation of a neutral allele. (See Exercise 3.9.)

Exercise 8.12. Let us use `wf.freq()` to see if the fixation time we just computed is well reflected in simulations. Take $s = 0.1, N = 10^3, 10^4, 10^5, 10^6$ with $p_0 = 1/2N$ and run `wf.freq()` for several times until the allele is fixed. Write down the time of fixation of the beneficial allele. How well do you results fit compared to (8.7)? □

The reason why there can still be variation on a chromosome after a sweep has happened is recombination. Think again of the A/a and the B/b locus. If the B occurs on a chromosome carrying an a-allele, the A-alleles will disappear completely unless a recombination event creates a chromosome with an A-allele and a B-allele.

In a two-locus two-allele model there are four possible gametic genotypes: ab, Ab, aB and AB. Recombination between an Ab and the aB chromosome can create an AB gamete. If this happens the a-allele will not go to fixation.

Exercise 8.13. Let us use `seqEvoHitchhiking()` from the `R`-package to see the effect of recombination. By calling e.g. `seqEvoHitchhiking(N=10,s=0.1,rho=0.1,u=0.1)` you see the sequence evolution during a selective sweep in a population of size 10, where the

recombination rate between the endpoints of the sequence is 0.1 and the mutation rate for the total sequence is 0.1. In particular, new mutations are created during the selective sweep.

1. Use the option `wait=-1` to see step by step, what changes during the sweep. Additionally, use `showFixed=TRUE` to see how many mutations become fixed during the sweep. (To see the same simulation twice, you can use `seed=1` or any other number.)

2. We argued that more variation is kept during the sweep if the recombination rate is larger. Set `rho`$= 0.01, 0.1, 1$ and record the number of segregating sites at the end of the sweep for 10 runs. Do you see that the level of sequence variation is higher for larger recombination rates?

□

The parameter r is the recombination rate between two sites. It is easily calculated by taking the recombination rate per nucleotide and multiplying by the distance between the two sites. $rho = r \cdot L$ (with r is recombination rate and L is the distance in nucleotides)

To understand the pattern of strong directional selection at the linked neutral A-locus we study genealogies at the end of the selective sweep. First consider the B/b locus. At the end of the sweep all individuals carry the B-allel; in contrast at the beginning only a single individual carries the beneficial B-allele. As a consequence, the B-allel of the only individual carrying the beneficial allele at the beginning of the sweep is a common ancestor to all beneficial alleles at the end of the sweep; see Figure 8.2(A).

Things change a bit at the neutral A/a-locus. Assume that we are given lines carrying an A/a-locus. At the end of the sweep we are sure that all are linked to a beneficial B allele. When we trace back a line it is either linked to B or b. By recombination events in the past it may change from B to b and back. We thus might draw lines changing backgrounds from the B to the b allele as in Figure 8.2(B).

Assume a recombination event occurs on the line we are considering when the B allele has frequency p. The probability that the recombinant carried a b allele rather than a B allele is $1-p$. Hence we say that a line changes background from B to b at rate $\rho(1-x)$. Using this description we can already compute the probability that a line changes background at some time in the sweep. The probability that it does *not* change background during some time dt when the frequency of B is p is given by $\exp(-\rho(1-p)dt)$. (This is a special case of the exponential distribuion with a non-constant rate; see Maths 1.3.) So, the probability that there is no change of backgrounds during the sweep is

$$\begin{aligned} q &= \exp\Big(-\rho\int_0^\tau (1-p_t)dt\Big) = \exp\Big(-\tfrac{2\rho}{s}\int_{p_0}^{1-p_0}\tfrac{1}{p}dp\Big) = \exp\Big(-\tfrac{2\rho}{s}\log\Big(\frac{1-p_0}{p_0}\Big)\Big) \\ &\approx \exp\Big(-\tfrac{2\rho}{s}\log(2N)\Big) \end{aligned} \tag{8.8}$$

and the probability that there is no such event is $1-q$. One can compute that back-recombinations from the b to the B-background can be ignored.

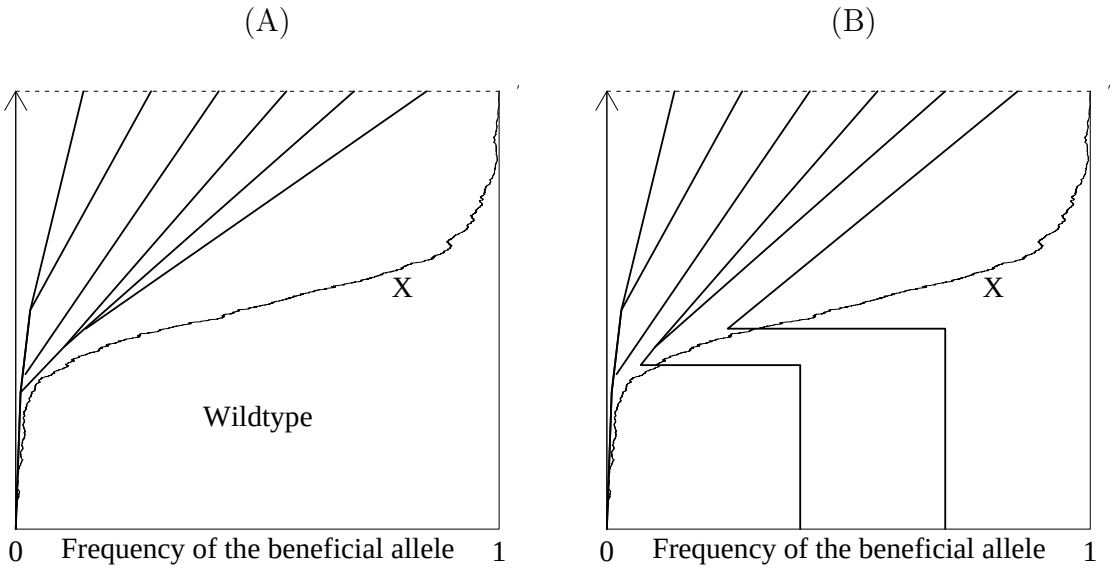


Figure 8.2: Genealogies at the selected and a linked neutral site. Lines above the curve are linked to a beneficial allele B; lines below the curve are linked to the wild-type b. (A) At the end of the sweep the genealogy at the selected site is almost star-like. (B) By recombination, a line of the A/a-locus can become linked to a wild-type b-allele.

To obtain the full genealogy of a sample of neutral loci we assume that the genealogy at the selected site is star-like. This is only an approximation because it might well be that some pairs of B-loci have a common ancestor during the sweep; see Figure 8.2(A).

Combining our calculations on the recombination of neutral loci and the genealogy at the selected site we have the following description: at the selected site there is a star-like genealogy. The length of each branch is $\frac{4}{s}\log(2N)$. The genealogy at the neutral loci is linked to this star-like genealogy. Namely, every branch is hit by a recombination event with probability $1-q$. Before the hitchhiking-event the genealogy is given by a neutral coalescent. The resulting genealogy is drawn in Figure 8.3.

Exercise 8.14. Consider the genealogy of Figure 8.3. Assuming that you see a typical genealogy at the end of a selective sweep, what kind of signature of a sweep do you expect to find in data? □

The most important signal of a selective sweep in data is the reduction of genetic variation. Next, let us make a finer analysis of the signature of a selective sweep in sequence data. Look at Figure 8.3 and assume that a mutation falls on the left branch before the hitchhiking event. Such a (neutral) mutation is carried by the founder of the hitchhiking event and therefore increases in frequency during the selective sweep. In the genealogy this is the same as saying that a lot of lines carry this neutral mutation. In other words, such a mutation occurs in high frequency in the sample. Such frequencies were the subject of the site frequency spectrum in 5.2.

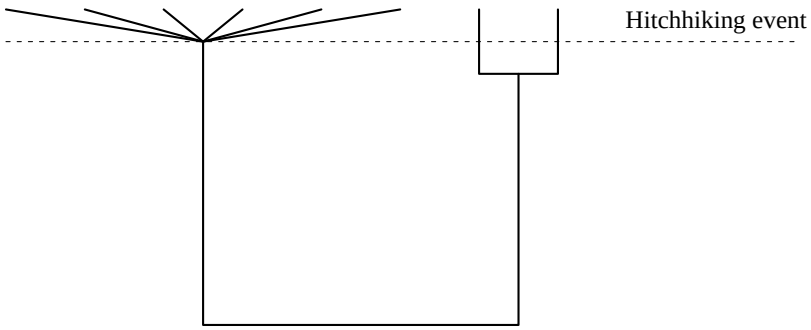


Figure 8.3: An approximate genealogy at the end of a selective sweep. The six leftmost lines are followed to the founder of the selective sweep while the two rightmosr lines are linked to a wild-type allel at the beginning of the sweep. Before the beginning of the sweep (at the lower part of the figure) lines coalesce as in a neutral coalescent.

Exercise 8.15. Let us use again `seqEvoHitchhiking()` from the `R`-package. To look at the frequency spectrum we use the option `sfs=TRUE`.

1. Consider a population of $N = 50$, say. Since the diversity pattern is so complicated, switch its display off using the option `seq=FALSE`. Change the values for rho from 0.001 to 1 and look at the evolution of the site frequency spectrum at the end of the sweep. For which values do you see an excess of high-frequency variants?

2. For $rho = 0.001$ you do almost never observe high frequency variants. Using the genealogy of Figure 8.3, can you explain why?

3. For $rho = 1$ or larger values of rho the frequency spectrum you observe looks almost like the neutral expectation (at least for the low frequency variatns). Why?

4. For which values of `rho` do you actually observe an excess of high-frequency variants?

□

9 Neutrality Tests

Up to now, we calculated different things from various models and compared our findings with data. But to be able to state, with some quantifiable certainty, that our data do not fit with the predictions, we need statistics. That means that it is not enough to describe phenomena, we need to quantify them and assign probabilities to observed data to see how probable they are under certain scenarios, such as neutrality or population expansion. Next we will deal with *neutrality tests*, which measure if data deviate from the expectations under a neutral model with constant population size. The first three tests we introduce, which all are called by the name of their test statistic, are Tajima's D, Fu and Li's D and Fay and Wu's H. These are based on data from a single population (plus one line of an outgroup to see which states are ancestral and which ones are derived). Then we are dealing with the HKA test and the McDonald-Kreitman test that both use data from two or more species or populations.

The statistical procedure is very general, only the used methods, i.e. what to compute from data, is unique to population genetics. Therefore we start with some general statistics.

9.1 Statistical inference

Let us start with a brief summary about statistical testing: Statistical testing starts by stating a *null hypothesis*, H_0. A *test statistic*, T, is chosen. If the value of T, calculated from the data, falls within a certain range of values called the *critical region*, $\mathcal{R}$, then the null hypothesis H_0 is rejected. The *size* of the test is $\alpha = \mathbf{P}[T \in \mathcal{R} \mid H_0]$. If the test statistic is scalar - (this is most often the case, and in this case you can say whether the statistic is smaller or bigger than a certain value) - then we can calculate a p-value, $p = \mathbf{P}[T \geq t \mid H_0]$ where the observed value of the test statistic is t. What did all of this mean? And more importantly, why is this the way we do statistical testing?

Firstly, the null hypothesis H_0. This has to be a mathematically explicit hypothesis. "There has been an effect of natural selection on my data" is not mathematically explicit, but unfortunately "There has not been an effect of natural selection on my data" is not sufficiently explicit either. Mathematical models are required for statistical testing. An example for a sufficiently explicit hypothesis would be:

> The population is in equilibrium, behaves like a Wright-Fisher model with constant population size $2N_e$. All mutations are neutral. The mutation rate is μ. There is no recombination.

With an H_0 like this many relevant parameters can be calculated and they can be compared with the observed data.

The test statistic T is a function of the data. It usually summarizes or condenses the data. There is a range of possible statistics that could be chosen. The aim is to choose one that contains the information we want, and ignores the information that we believe is irrelevant, as far as is possible. For the test to work at all, it is essential to know the distribution of T when H_0 is true, written $\mathbf{P}[T \mid H_0]$. Sometimes this distribution can be

calculated analytically but for more complex tests it may have to be estimated by computer simulation.

The size of the test, $\alpha = \mathbf{P}[T \in \mathcal{R} \mid H_0]$, is the probability that the null hypothesis will be rejected when it is in fact true. This is a *false positive* error, a.k.a. a type I error. We can control the chance of such an error occurring, by our choice of either T or $\mathcal{R}$. Note that some supposedly authoritative sources say that the p-value is the probability of making a type I error. *This is not true!* Only the size of the test, if determined before the data are inspected, has this property.

An interpretation of a p-value is the probability of observing data like the data that was observed *or more extreme*, given the null hypothesis. These p-values can only be calculated for *scalar* test statistics. This is because we need to define an order, so that we can say which data are more extreme than others.

The other type of error is a *false negative*, a.k.a. a type II error, which is a failure to reject the null hypothesis when it is in fact wrong. We cannot control the chance of these errors occurring; they depend on what alternate hypothesis is true instead of H_0. If an alternative hypothesis H_1 is in fact true, then the power of the test for H_1 is $\mathbf{P}[T \in \mathcal{R} \mid H_1]$, which is determined by the choice of T and the critical region $\mathcal{R}$. High power is desirable. Therefore, in order to design a good test it is important to have a good idea about which alternative hypotheses could be true. For genetic data there are usually an infinity of possible choices for T and $\mathcal{R}$, so some kind of biological insight is important.

Exercise 9.1. Assume you are given two datasets, `data1` and `data2`. You perform a test of neutrality on both of them. For `data1` you get a significant result (with $\alpha = 5\%$) and for `data2` a non-significant one. Which of the following conclusions can you draw?

- The dataset `data1` does not stem from a neutral model of constant size.
- The dataset `data2` stems from a neutral model of constant size.

Which error is involved in these two conclusions? Which error is controlled by the size of the test? □

Example: Fisher's exact test

Let us look at an example, taken from Sokal and Rohlf (1994). We will be dealing with *Fisher's exact test* which will be of relevance in this section. This test uses as data a 2×2-contingency table. This is a table of the form given in Figure 9.1. Here *acacia trees* were studied and whether or not they are invaded by ant colonies. The aim of the study was to find out whether species A is more often invaded by ant colonies than species B.

From species A, 13 out of 15 trees were invaded, but only 3 out of 13 from species B. So it certainly looks as if A is more often invaded. But we would like to know whether this is statistically significant. Now, you know that in the study there is a total of 15 trees from species A and 13 from species B. and you know that 16 trees are invaded by ants and 12 are not. Using only this data, and if you would assume that both species are

Acacia species	Not invaded	Invaded	Total
A	2	13	15
B	10	3	13
Total	12	16	28

Figure 9.1: Example of data to use Fisher's exact test.

equally likely to be invaded you would expect that $16 \cdot \frac{15}{28} \approx 8.57$ trees of species A would be invaded. This value is the expected value under the assumption that the species of a tree and whether or not it is invaded are two independent things.

You already know of the χ^2-test which can also be used in this case. To make the χ^2-test all you need is your data in the contingency table and the expectations like the one we just computed. Then, you calculate, as on page 58,

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \\ &= \frac{(2 - \frac{15}{28}12)^2}{\frac{15}{28}12} + \frac{(13 - \frac{15}{28}16)^2}{\frac{15}{28}16} + \frac{(10 - \frac{13}{28}12)^2}{\frac{13}{28}12} + \frac{(3 - \frac{13}{28}16)^2}{\frac{13}{28}16} \approx 11.4991\end{aligned}$$

Usually, you now say that the statistic you just computed is χ^2-distributed with $(\text{rows} - 1)(\text{lines} - 1) = 1$ degree of freedom. You can then look up in a table of the χ^2-distribution with 1 degree of freedom that a value of 11.4991 or larger only appears with probability $p = 0.0006963$ which is then also the p-value. However, all of this relies on the χ^2-distribution of the statistic we computed. And in this case the statistic is not exactly χ^2-distributed, because our data are discrete and not continuous, as they would have to be in order to be χ^2-distributed. There are corrections for this, but here we will use a different method: Fisher's exact test.

The test is called *exact* because the distribution of the test statistic is exact and not only approximate as it would be for the χ^2-test. Fisher's exact test relies on computing the probability of obtaining the observed data given the marginals of the contingency table. To compute these the number of ways to put 28 balls into four urns such that all marginals are correct is

$$\binom{28}{15}\binom{28}{16}.$$

To calculate the number of ways to obtain not only the marginals but the numbers in the four cells assume you must lay 28 balls in a row, where 2 have color a, 13 have color b, 10 have color c and 3 have color d. The color a balls can be put on the 28 sites in $\binom{28}{2}$ ways. There are 26 positions remaining. Next choose 13 positions for the balls of color b, which

give $\binom{26}{13}$ possibilities. Afterwards, for color c you have $\binom{13}{10}$ possibilities. The balls of color d must then be put in the remaining positions. This totally gives

$$\binom{28}{2}\binom{26}{13}\binom{13}{10} = \frac{28!}{2!13!10!3!}.$$

Let us assume we have data a, b, c, d instead of the given numbers. Then, as the probability of the observed data must be the number of ways of obtaining it divided by the number of all possibilities, we have

$$\mathbf{P}[(a,b,c,d)] = \frac{\frac{n!}{a!b!c!d!}}{\binom{n}{a+b}\binom{n}{a+c}} = \frac{(a+b)!(a+c)!(b+c)!(b+d)!}{a!b!c!d!n!}. \tag{9.1}$$

So in our case we have

$$\mathbf{P}[(2,13,10,3)] = \frac{15!13!12!16!}{28!2!13!10!3!} = 0.00098712.$$

which is the probability of finding the contingency table that we had. However, the p-value was defined as the probability that, given the null-hypothesis, the data are at least as extreme as the observed data. The data would be more extreme if the data would like one of the tables given in Figure 9.2. Note however, that the marginals of the contingency table are fixed here. We only check independence of invasion and species given these marginals.

Using these more extreme cases we can calculate the p-value, by adding up the probabilities for all the more extreme cases. It turns out to be 0.00162. This means that the test is highly significant and the hypothesis, that the invasion of ants is independent of the species can be rejected. All about Fisher's exact test is summarized in Figure 9.3.

The easiest way to perform tests on contigency tables is by using a web-based calculator. You can find a χ^2-calculator e.g. at `http://schnoodles.com/cgi-bin/web_chi.cgi`, one for Fisher's exact test is found at `http://www.matforsk.no/ola/fisher.htm`. Also programs like `DNASP` (use `Tools->Tests of Independence: 2x2 table`) and of course any statistical package like `R` can do such tests.

Exercise 9.2. A plant ecologist samples 100 trees of a rare species from a 400-square-kilometer area. His records for each tree whether or not it is rooted in serpentine soils and whether its leaves are pubescent or smooth. The data he collected in Figure 9.4.

Use

1. a χ^2-test
2. Fisher's exact test

to assess whether the kind of soil and the kind of leaves are independent. Compute the p-values in each case? Interpret your results. □

Acacia species	Not invaded	Invaded	Total
A	1	14	15
B	11	2	13
Total	12	16	28

Acacia species	Not invaded	Invaded	Total
A	0	15	15
B	12	1	13
Total	12	16	28

Acacia species	Not invaded	Invaded	Total
A	11	4	15
B	1	12	13
Total	12	16	28

Acacia species	Not invaded	Invaded	Total
A	12	3	15
B	0	13	13
Total	12	16	28

Figure 9.2: More extreme cases for Fisher's exact test

9.2 Tajima's D

Recall that, if the neutral theory and the infinite sites model hold, there are a number of different unbiased estimators of $\theta = 4N_e\mu$. These include the estimator $\widehat{\theta}_S$ (see (2.7)) where S is the number of segregating sites in a sample of n sequences. A second was given in (2.6) to be $\widehat{\theta}_\pi$ which is the mean pairwise difference between the sequences in the sample. Both these estimators are unbiased but they use different information in the sample. This motivated Tajima (1989) to propose his d statistic, which is defined as

$$d := \widehat{\theta}_\pi - \widehat{\theta}_S. \tag{9.2}$$

Since both estimators are unbiased, for neutral/infinite sites model $\mathbf{E}[d] = 0$. However, because the two statistics have different sensitivities to deviations from the neutral model, d contains information.

Exercise 9.3. Take the data from Exercise 5.1. Compute d in this case either by hand or using `R`. □

We know that the expectation of d is 0, but in order to use it as a test statistic we need to know also the variance. As Tajima (1989) showed, the variance can be estimated by (we don't give the derivation because it is too long)

$$\widehat{\mathbf{Var}}[\widehat{\theta}_\pi - \widehat{\theta}_S] = \frac{c_1}{a_1}S + \frac{c_2}{a_1^2 + a_2}S(S-1)$$

Fisher's exact test

checks for independence in a 2×2 contingency table.

Data

a, b, c, d as given below; their distribution is any distribution for which $a+b$, $a+c$, $b+c$ and $b+d$ are fixed

	Case 1	Case 2	Total
Case A	a	b	$a+b$
Case B	c	d	$c+d$
Total	$a+c$	$b+d$	n

Null-hypothesis

a, b, c, d are distributed independently on the table, leaving $a+b, a+c, b+c, b+d$ fixed,

$$\mathbf{P}[(a,b,c,d)|a{+}b, a{+}c, b{+}c, b{+}d, n] = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

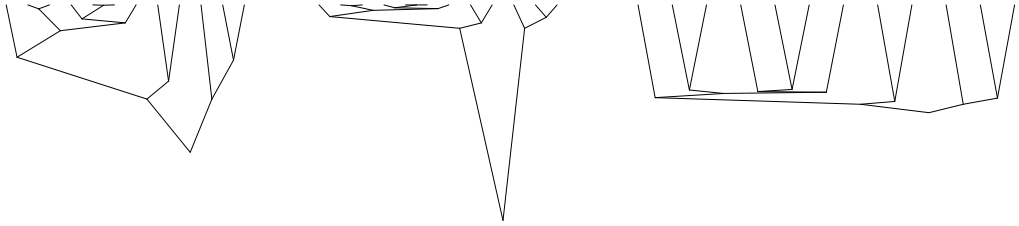
p-value

$$p = \sum_{\substack{a,b,c,d \text{ at least} \\ \text{as extreme as data}}} \mathbf{P}[(a,b,c,d)|a+b, a+c, b+c, b+d, n].$$

Figure 9.3: Schematic use of Fisher's exact test

Leaf form	Serpentine soil	No serpentine soil	Total
Pubescent	19	18	37
Smooth	17	46	63
Total	36	64	100

Figure 9.4: Data for Exercise 9.2



Neutral	Balancing selection	Recent sweep
$D = 0$	$D > 0$	$D < 0$

Figure 9.5: Genealogies under different forms of selection.

with

$$c_1 = b_1 - \frac{1}{a_1}, \qquad c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}, \tag{9.3}$$

$$b_1 = \frac{n+1}{3(n-1)}, \qquad b_2 = \frac{2(n^2+n+3)}{9n(n-1)}, \tag{9.4}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}, \qquad a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}. \tag{9.5}$$

Using this he defined the test statistic

$$D = \frac{\widehat{\theta}_\pi - \widehat{\theta}_S}{\sqrt{\widehat{\mathbf{Var}}[\widehat{\theta}_\pi - \widehat{\theta}_S]}}. \tag{9.6}$$

Tajima's D statistic is probably the most widely used test of neutrality. It has the appealing property that its mean is approximately 0 and its variance approximately 1. Also, for large sample sizes it is approximately normally distributed under neutrality, and more generally it is approximately β-distributed. However, it is still preferable not to use the approximate distribution, but to get the real one by doing computer simulations.

What values of Tajima's D do we expect when there is a deviation from the neutral model? The statistic reflects the shape of the genealogy. Consider the genealogies of equal total length shown in Figure 9.5.

Keeping the total length constant means that $\mathbf{E}[\widehat{\theta}_S]$ is constant, because a mutation anywhere on the tree causes a segregating site. We can transform between genealogies by moving one node up and another node down, without changing the total length. For example, moving the root down one unit increases the pairwise distance by two units for $9 \cdot 3 = 27$ pairwise comparisons, but moving any point up by two units increases the pairwise distance by four units for a smaller number of pairwise comparisons. The net effect is to change the expected pairwise difference is thus positive. This illustrates why genealogies which have a deep bifurcation tend to have positive values of Tajima's D. For

the same reasons, genealogies that are approximately star-like tend to have negative values of Tajima's D.

We expect a selective sweep to cause a star shaped genealogy and balancing selection to cause a deeply bifurcated genealogy. Purifying selection against recurrent deleterious mutations causes a small reduction in D because deleterious alleles are at low frequency on short branches.

Exercise 9.4. The last paragraph gave a heuristic view about genealogical trees under different selective forces. Explain in your own words why

- selective sweeps cause star like genealogies,
- balancing selection leads to deeply bifurcated trees.

Which kind of tree topologies from Figure 9.5 would you expect in a substructured population? □

Conditional Testing of Tajima's D

Recall that for the denominator of D the variance of d was estimated using an estimator of θ, $\widehat{\theta}$. So when we calculate a p-value for Tajima's D we use exactly this estimator. Implicitly we assume here that the estimator gives us the correct value. So, when we want to find out the distribution of Tajima's D and p-values, that follow from the distribution, we want to find

$$p = \mathbf{P}[D < d|\widehat{\theta}, \text{neutrality}]. \tag{9.7}$$

With modern computers and fast coalescent simulations, there is really no reason to approximate the sampling distribution of D with a normal or beta distribution. Instead, the exact sampling distribution should be estimated by simulation.

Exercise 9.5. `DNASP` can do these coalescent simulations for you. Open the file `hla-b.nex`. Here you see 50 lines from the human *hla* region. This locus is supposed to be under balancing selection, so Tajima's D should be positive. Always use the complete dataset in this exercise.

1. Does Tajima's D-test give a significant result for the region as a whole?
2. Do a sliding window analysis to see how Tajima's D behaves in the region of consideration. Here `DNASP` uses a predefined distribution of Tajima's D, probably a normal distribution.
3. You will see that there is a region where Tajima's D tends to be positive but not significantly so (i.e. with a p-value above 5%). Use `Tools->Coalescent Simulations` to compute the critical region of Tajima's D test, given the overall level of polymorphism, e.g. given by $\widehat{\theta}_S$ or $\widehat{\theta}_\pi$. Do your simulation results support the hypothesis of balancing selection better than your results in 1?

□

Some more recent papers have instead calculated p-values using the *conditional* sampling distribution of D, obtained by conditioning on the observed number of segregating sites, i.e. $S = s$. The reason for doing this is that no assumption has to be made that an estimated θ-value is correct. The number of segregating sites is just as it is. So these papers compute

$$p = \mathbf{P}[D < d | S = s, \text{neutrality}]. \tag{9.8}$$

These two ways of tests using Tajima's D produce different p-values and different critical regions. For example, suppose we know $\theta = 1$ or we estimated $\widehat{\theta} = 1$ and assume this is correct. Then by generating 10^5 random samples from the distribution of D, for $n = 30$, we find that the 0.025 and 0.975 quantiles -1.585 and 1.969 respectively.

Thus $\mathcal{R} = \{D \leq -1.585 \text{ or } D \geq 1.969\}$ is a rejection region for a size 0.05 test, i.e. a test that guarantees to make false rejection errors with a frequency at worst 5% in a long series of trials. However, we conditioned here on having found the correct θ. When we condition on the number of segregating sites s, the frequency of making a type I error by rejecting neutrality using rejection region R is not 5%, as shown in Figure 9.6.

Exercise 9.6. In `Tools->Coalescent Simulations` `DNASP` offers a coalescent simulation interface. Let us use this to see the differences between conditioning on $\widehat{\theta} = \theta$ and on $S = s$.

1. Make 10000 coalescent simulations to check whether you also obtain the critical region R. However it is possible that the values `DNASP` gives you a deviate from the above ones though . How do you interpret this?

2. `DNASP` offers you to simulate `Given Theta` and `Given Segregating Sites` which is exactly the distinction we also draw here. Using the coalescent interface can you also obtain the values given in Figure 9.6?

□

9.3 Fu and Li's D

Another test that is directly based on the coalescent was proposed by Fu and Li (1993). Their test statistic is based on the fact that singletons, i.e. polymorphisms that only affect one individual in a sample, play a special role for different population histories. From the frequency spectrum (see Section 5) we know that

$$\mathbf{E}[S_i] = \frac{\theta}{i},$$

s	$\mathbf{P}[D \in R]$	bias
1	0%	less type I
4	6.22%	more type I
7	6.93%	more type I

Figure 9.6: Assume a critical region $\mathcal{R} = \{D \leq -1.585 \text{ or } D \geq 1.969\}$. This should fit to a test of site 5% where $n = 30$ and $\theta = 1$. Conditioning on the number of segregating sites the frequencies of type I errors change. A total of 10^5 unconditional trials were simulated.

where S_i is the number of segregating sites that affect i individuals in the sample. Using this, we can e.g. conclude that

$$\mathbf{E}[S] = \sum_{i=1}^{n-1} \mathbf{E}[S_i] = \theta \sum_{i=1}^{n-1} \frac{1}{i},$$

which gives a better understanding of the estimator $\widehat{\theta}_S$. We can also use it to predict the number of singletons: $\mathbf{E}[S_1] = \theta$, this gives us a new unbiased estimator of θ:

$$\widehat{\theta}_{S_1} = S_1$$

With their test statistic, Fu and Li (1993) compared the level of polymorphism on external branches, which are the singleton mutations, with the level of polymorphism on internal branches, i.e. all other mutations. As

$$\sum_{i=2}^{n-1} \mathbf{E}[S_i] = \theta \sum_{i=2}^{n-1} \frac{1}{i},$$

another unbiased estimator of θ is

$$\widehat{\theta}_{S_{>1}} = \frac{S_{>1}}{\sum_{i=2}^{n-1} \frac{1}{i}}$$

where $S_{>1}$ are all segregating sites that affect more than one individual. In the same spirit as Tajima (1989), Fu and Li (1993) proposed the statistic

$$d = \widehat{\theta}_{S_{>1}} - \widehat{\theta}_{S_1}.$$

Exercise 9.7. Again use data from Exercise 5.1 (plus the outgroup of Exercise 5.3). Compute d in this case either by hand or using `R`. □

Again - in order to make it a better test statistic - they calculated the variance of d and found out that a good estimator of this is

$$\widehat{\mathbf{Var}}[\widehat{\theta}_{S_{>1}} - \widehat{\theta}_{S_1}] = c_1 S + c_2 S^2$$

with

$$\begin{aligned} c_1 &= 1 + \frac{a_1^2}{a_2 + a_1^2}\Big(b - \frac{n+1}{n-1}\Big), \\ c_2 &= a_n - 1 - c_1, \\ b &= \frac{2na_n - 4(n-1)}{(n-1)(n-2)}, \end{aligned}$$

and a_1, a_2 from (9.3).

Analogously to Tajima's D statistic Fu and Li (1993) proposed the test statistic

$$D = \frac{\hat{\theta}_{S_{>1}} - \hat{\theta}_{S_1}}{\sqrt{\widehat{\mathbf{Var}}[\hat{\theta}_i - \hat{\theta}_e]}}.$$

The interpretation for selective models can again be read from Figure 9.5. When the tree looks like the one in the middle θ_i will be small and D in this case negative. When the genealogical tree looks like the right side then $\hat{\theta}_{S_{>1}}$ will be far too small and so D will positive.

Exercise 9.8. Use the same dataset as in Exercise 9.5. Does Fu and Li's D suggest that balancing selection is going on at the human HLA locus? Again answer this question using the dataset as a whole, for a sliding window analysis and using coalescent simulations. □

Exercise 9.9. Obviously Fu and Li's D and Tajima's D use different information of the data. Can you draw a genealogical tree (with mutations on the tree) for the case that

- Tajima's D is negative and Fu and Li's D is approximately 0?
- Fu and Li's D is positive and Tajima's D is approximately 0?

□

9.4 Fay and Wu's *H*

Tests based on polymorphism data only, are easily confounded by demography and by selection at linked sites. These problems can be addressed to some extent by combining divergence and polymorphism data, and by combining data from several loci at once.

As we know from Section 8, the genealogy at loci closely but not completely linked to a site at which a selective sweep has occurred, there is a tendency for the genealogies to resemble the one in Figure 8.3.

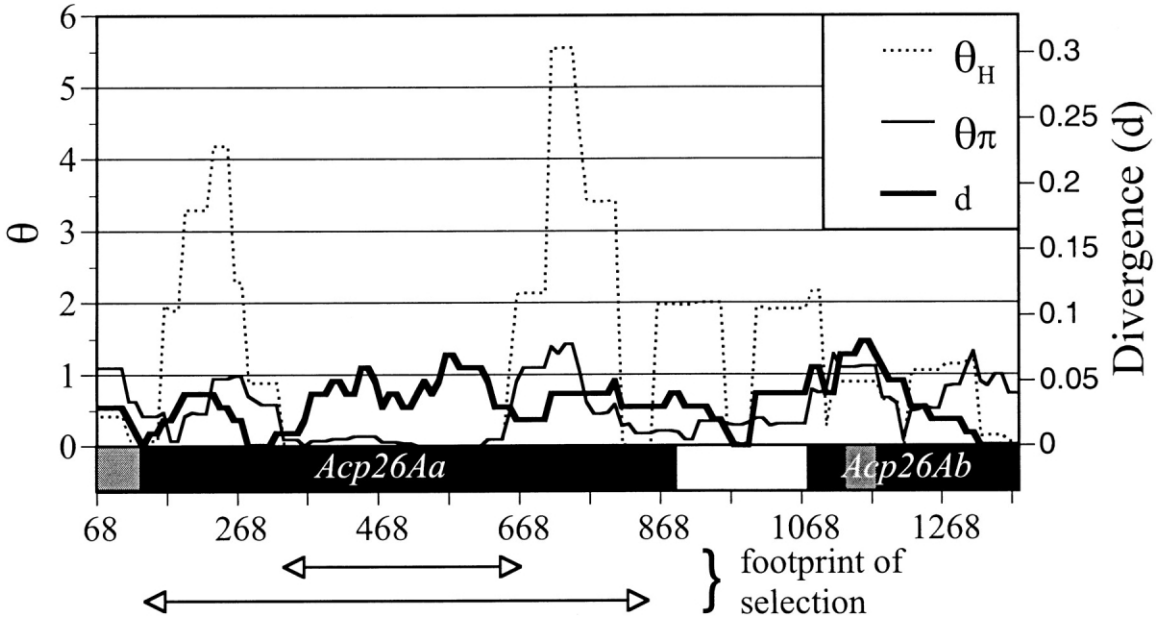


Figure 9.7: Sliding window analysis of the accessory gland protein gene by Fay and Wu (2000).

Especially, we saw in Exercise Ex:hitchSFS that hitchhiking give rise to *high frequency derived alleles*. The statistic

$$\widehat{\theta}_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \tag{9.9}$$

is particularly sensitive to the presence of such variants, and therefore suggests the test statistic

$$H = \widehat{\theta}_T - \widehat{\theta}_H \tag{9.10}$$

Fay and Wu did not provide an analytical calculation of the variance for this statistic and therefore p-values have to be estimated by simulation. In both simulated and real data for which divergence data have indicated positive selection, Fay and Wu (2000) found smaller p-values for H-tests than for Tajima's D-tests.

Particularly spectacular are the large peaks in $\widehat{\theta}_H$ for the Drosophila accessory gland protein gene (Figure 9.7); perhaps two regions where there could have been incomplete hitchhiking, one on either side of a $\sim$ 350bp region where almost all variability has been eliminated.

9.5 The HKA Test

The HKA test that was introduced by Hudson *et al.* (1987), uses polymorphism and divergence data from two or more loci, and tests for loci that have an unusual pattern of molecular variation. Here we will consider the case of L loci and polymorphism data from two species A and B.

The idea and the model

The main idea of the HKA test is that under neutrality you would expect the same ratio between polymorphism within and divergence between species in the L loci under observation.

The test assumes that the polymorphism data is taken from two species with a fixed effective population size of $2N_e$ and $2N_ef$ haploids. These species are assumed to have diverged from each other some time T ago, where T is measured in units of $1/2N_e$ generations. The ancestral species is assumed to have had a fixed effective population size of $2N_e\frac{1+f}{2}$ diploids. Further assumptions are:

1. at each locus an infinite sites model with selectively neutral mutations holds
2. the mutation rate at locus i is μ_i $(i = 1, \ldots, L)$
3. there is no recombination within the loci
4. there is free recombination between the loci, i.e. loci are unlinked

Parameters and data

On the data side we have

$S_1^A, \ldots, S_L^A$	number of segregating sites in species A at loci $1, \ldots, L$
$S_1^B, \ldots, S_L^B$	number of segregating sites in species B at loci $1, \ldots, L$
$D_1, \ldots, D_L$	divergence between two randomly picked lines from species A and B at loci $1, \ldots, L$.

So there are $3L$ numbers that we observe. On the parameter side we have:

T	time since the split of the two species
f	ratio of the two population sizes of species B and A
$\theta_1, \ldots, \theta_L$	scaled mutation rate at locis $1, \ldots, L$

So there are $L + 2$ model parameters. As long as $L \geq 2$, i.e. when data of two or more loci is available there are more observations than parameters. This means that we can test the model.

Exercise 9.10. Why does it only make sense to test the model if the data consists of more numbers than the number of model parameters? □

Estimation of model parameters

First of all we have to estimate the model parameters. This can be done by different means. As the estimator $\widehat{\theta}_S$ is unbiased for one locus, we can calculate that

$$\sum_{i=1}^{L} S_i^A = a_{n_A} \sum_{i=1}^{L} \widehat{\theta}_i, \tag{9.11}$$

$$\sum_{i=1}^{L} S_i^B = a_{n_B} f \sum_{i=1}^{L} \widehat{\theta}_i, \tag{9.12}$$

where

$$a_n := \sum_{i=1}^{n-1} \frac{1}{i},$$

and n_A is the number of sequences we have for species A. Divergence of two randomly picked lines from the two populations is on average $2N_e(T + \frac{1+f}{2})$ generations. So, we can estimate

$$\sum_{i=1}^{L} D_i = \big(\widehat{T} + \tfrac{1+\widehat{f}}{2}\big)\widehat{\theta}_i. \tag{9.13}$$

For each single locus $i = 1, \ldots, L$ we have polymorphism in A, polymorphism in B and divergence which adds up to

$$S_i^A + S_i^B + D_i = \widehat{\theta}_i\big(\widehat{T} + \tfrac{1+\widehat{f}}{2} + a_{n_a} + a_{n_B}\big). \tag{9.14}$$

Using these equations we would have $L + 3$ equations for $L + 2$ unknowns, which are the model parameters. So it is not guaranteed that we find a solution. But we can try to find them by a least squares approach. Assume you take some combination of $\widehat{t}, \widehat{f}, \widehat{\theta}_1, \ldots, \widehat{\theta}_L$. When you plug them into the right sides of (9.11)-(9.14) these numbers produce some numbers for the left sides for these equations. They are least squares estimators if and only if the sum of squares of differences of those produced left sides to the real ones, given by the left sides of (9.11)-(9.14), is minimal.

The test statistic

Ultimately we want to test if our model fits the observations. So far we have estimated the model parameters. The HKA test now makes the assumption that

- the random numbers $S_1^A, \ldots, S_L^A, S_1^B, \ldots, S_L^B, D_1, \ldots, D_L$ are independent and normally distributed.

While for large data sets it is possible that approximately the normal distribution holds true. The independence is certainly false, e.g. because the divergence uses the same segregating sites as the polymorphism. The HKA test has been criticized for this. Let us see what happens under this assumption.

Next we need two properties of probability distributions.

Maths 9.1. *When X is normally distributed, then*

$$Y := \frac{X - \mathbf{E}[X]}{\mathbf{Var}[X]}$$

is also normally distributed with $\mathbf{E}[Y] = 0, \mathbf{Var}[Y] = 1$.

Maths 9.2. *Let $X_1, \ldots, X_n$ be independent, normally distributed random variables with $\mathbf{E}[X_i] = 0, \mathbf{Var}[X_i] = 1$ $(i = 1, \ldots, n)$. Then the distribution of*

$$Z = X_1^2 + \ldots + X_n^2 \tag{9.15}$$

is $\chi^2(n)$ distributed. Here n denotes the number of degrees of freedom.

When the above assumption holds at least approximately we now see that

$$\frac{S_i^A - \mathbf{E}[S_i^A]}{\sqrt{\mathbf{Var}[S_i^A]}}$$

is approximately normally distributed and

$$\sum_{i=1}^{L} \frac{S_i^A - \mathbf{E}[S_i^A]}{\mathbf{Var}[S_i^A]} + \sum_{i=1}^{L} \frac{S_i^B - \mathbf{E}[S_i^B]}{\mathbf{Var}[S_i^B]} + \sum_{i=1}^{L} \frac{D_i - \mathbf{E}[D_i]}{\mathbf{Var}[D_i]}$$

is approximately χ^2 distributed. But as we do not know $\mathbf{E}[S_i^A]$ and $\mathbf{Var}[S_i^A]$ we have to estimate them before we can use this. This is easy for the expectation, but for the variance we have to compute something.

Assume a coalescent tree for n individuals with mutation rate μ. Let L be the length of the whole tree and T_i be the time the tree spends with i lines. As usual S is the number of segregating sites. Then

$$\begin{aligned}
\mathbf{E}[S(S-1)] &= \int_0^\infty \mathbf{E}[S(S-1)|L = \ell]\mathbf{P}[L \in d\ell] \\
&= \int_0^\infty \sum_{s=0}^{\infty} s(s-1)e^{-\ell\mu}\frac{(\ell\mu)^s}{s!}\mathbf{P}[L \in d\ell] \\
&= \int_0^\infty (\ell\mu)^2 \sum_{s=2}^{\infty} e^{-\ell\mu}\frac{(\ell\mu)^s}{s!}\mathbf{P}[L \in d\ell] \\
&= \mu^2 \int_0^\infty \ell^2 \mathbf{P}[L \in d\ell] = \mu^2 \mathbf{E}[L^2] = \mu^2\big(\mathbf{Var}[L] + (\mathbf{E}[L])^2\big).
\end{aligned}$$

So we have to compute the variance of L. As we calculated in Maths 1.3 the variance of an exponentially distributed variable we can continue

$$\mathbf{Var}[L] = \mathbf{Var}\Big[\sum_{i=2}^{n} iT_i\Big] = \sum_{i=2}^{n} i^2 \mathbf{Var}[T_i] = \sum_{i=2}^{n} i^2 \Big(\frac{2N}{\binom{i}{2}}\Big)^2 = (2N)^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

i	locus	s_i	d_i
1	*Adh*	20	16
2	5' flanking region	30	78
3	*Adh-dup*	13	50

Figure 9.8: Data from 11 *D. melanogaster* and 1 *D. simulans* sequences

and so

$$\mathbf{Var}[S] = \mathbf{E}[S] + \mathbf{E}[S(S-1)] - (\mathbf{E}[S])^2 = \theta a_1 + \theta^2(a_2 + a_1^2 - a_1^2) = \theta a_1 + \theta^2 a_2. \quad (9.16)$$

Using this we can estimate, for the ith locus in species A

$$\widehat{\mathbf{Var}}[S_i^A] = \widehat{\theta}_i a_1 + \widehat{\theta}_i^2 a_2$$

and in species B

$$\widehat{\mathbf{Var}}[S_i^B] = \widehat{\theta}_i \widehat{f} a_1 + (\widehat{\theta}_i \widehat{f})^2 a_2.$$

The same calculation also works for divergence and in this case

$$\widehat{\mathbf{Var}}[D_i] = \widehat{\theta}_i\big(\widehat{T} + \frac{1+\widehat{f}}{2}\big) + \big(\widehat{\theta}_i\big(\widehat{T} + \frac{1+\widehat{f}}{2}\big)\big)^2.$$

Now we can use the test statistic

$$\chi^2 = \sum_{i=1}^{L} \frac{S_i^A - \widehat{\mathbf{E}}[S_i^A]}{\widehat{\mathbf{Var}}[S_i^A]} + \sum_{i=1}^{L} \frac{S_i^B - \widehat{\mathbf{E}}[S_i^B]}{\widehat{\mathbf{Var}}[S_i^B]} + \sum_{i=1}^{L} \frac{D_i - \widehat{\mathbf{E}}[D_i]}{\widehat{\mathbf{Var}}[D_i]}. \quad (9.17)$$

Each estimated parameter reduces the degree of freedom of this test statistic by 1 and so χ^2 has a χ^2-distribution with $3L - (L+2) = 2L - 2$ degrees of freedom. Now we have a test statistic and also know approximately its distribution, so we can calculate critical regions and p-values.

Example

We can apply the HKA test to data for the *Adh* gene, its 5' flanking region and an ancient duplicate gene *Adh-dup*, shown in Figure 9.8. The data are from the paper in which the HKA test was introduced Hudson *et al.* (1987).

Here we have 11 lines from *D. melanogaster* but only one from *D. simulans*. The consequence is the for *simulans* we do not have polymorphism data. As our data is less we need also less model parameters. We do this by assuming that $f = 0$. This means that we assume that *D. simulans* and *D. Melanogaster* have emerged from one species that had the size of *D. melanogaster* today. The *D. simulans* has split as a small fraction of this ancestral species. This is illustrated in Figure 9.9.

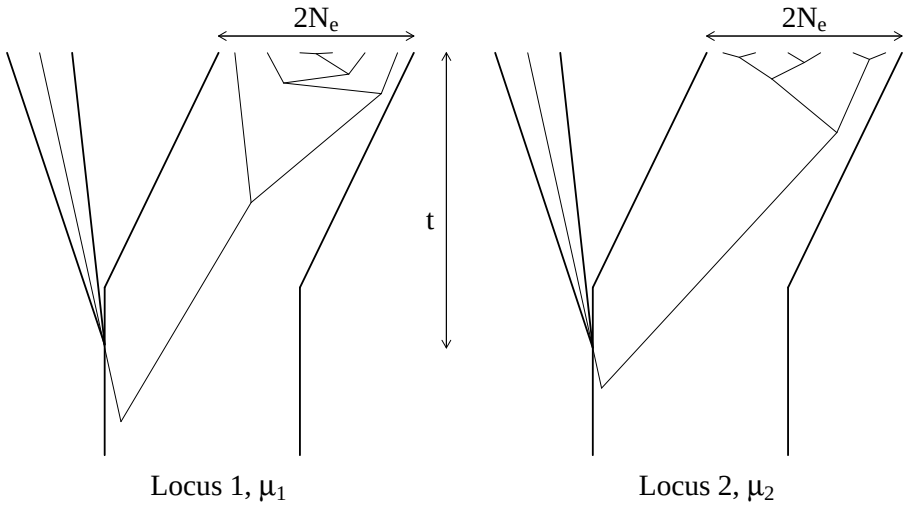


Figure 9.9: Model assumed by the HKA test in the special case $f = 0$ and polymorphism data only from one species.

The only significant pairwise comparison is between *Adh* and *Adh-dup*, which has $\chi^2 = 4.574$ and $p = 0.0325$ assuming the χ^2 approximation. For this test the expected quantities were ($\widehat{\mathbf{E}}[S_1] = 12.0$, $\widehat{\mathbf{E}}[S_3] = 21.0$, $\widehat{\mathbf{E}}[D_1] = 24.0$, $\widehat{\mathbf{E}}[D_3] = 42.0$). Comparing these with the observed values suggests that the *Adh* gene has unusually high level of polymorphism or an unusually low divergence, or *vice versa* for *Adh-dup*. This has been interpreted as evidence of balancing selection acting on the *Adh* fast—slow polymorphism.

Exercise 9.11. Two versions of the HKA test are implemented in `DNASP`. However it can only deal with two loci and always sets $f = 0$ in the above analysis. For the first (`Analysis->HKA, Hudson, Kreitman, Aguade Test`) you have to define the two loci in your data set. The second version (`Tools->HKA Test (Direct Mode)`) only needs numbers of segregating sites and divergence between two species as input. Here, it is assumed that only one line from species B is available, so we find ourselves in the regime of the above example.

1. Can you reproduce the numbers of the above example, e.g. $\chi^2 = 4.574$?
2. Is the comparison *Adh* and *Adh-dup* really the only one that gives a significant result?

□

Note that the parameter estimates obtained using equations (9.11)-(9.14) are not the values that minimize the test statistic. The least squares procedure minimizes the sum

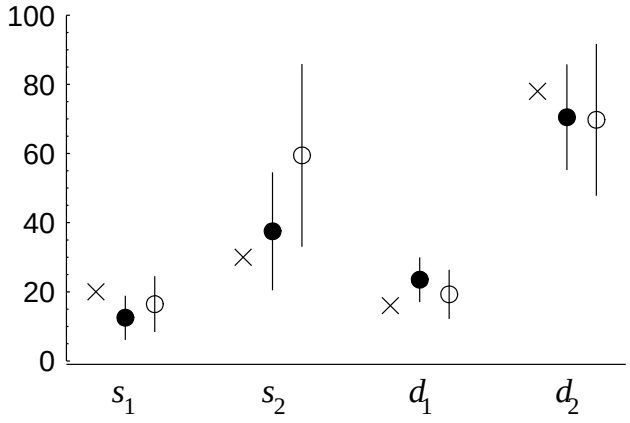


Figure 9.10: Data (×) for HKA test for *Adh* locus versus its 5' flanking region. Expectations ± 1 s.d. are shown for each quantity, calculated using either the parameter estimates obtained by equations (9.11)-(9.14) (closed symbols) or the estimates that minimize the test statistic of equation (9.17) (open symbols).

of the numerators in equations (9.11)-(9.14) without regard to differences in the expected variance of each quantity. This is illustrated in Figure 9.10 for the comparison between *Adh* and its 5' flanking region (table 9.8). Parameter estimates obtained by equations (9.11)(9.14) are $t = 4.5$, $\theta_1 = 4.3$, $\theta_2 = 12.8$ and give $\chi^2 = 3.2$. However, the minimum $\chi^2_{\text{Min}} = 1.79$ is obtained when $t = 2.4$, $\theta_1 = 5.6$, $\theta_2 = 20.3$.

Estimating parameters by minimizing χ^2 produces smaller values of the test statistic X^2, and hence larger P-values if it is assumed that the null distribution of χ^2 is χ^2 distributed with 1 degree of freedom. The HKA test is quite a robust test. Many deviations from the model, for example linkage between the loci or a population bottleneck in the past, generate correlations between the genealogies at the two loci and therefore reduce the variance of the test statistic, making the test conservative.

Exercise 9.12. One property of an estimator is consistency. That means that it gets better when more data is available. Assume an estimator $\widehat{\bullet}^{(n)}$ of $\bullet$ is based on n data. Consistency means that

$$\mathbf{Var}[\,\widehat{\bullet}^{(n)}\,] \xrightarrow{n\to\infty} 0.$$

You know the estimator $\widehat{\theta}_S$ which is only based on S. We could also call it $\widehat{\theta}_S^{(n)}$ when the sequences of n individuals are available. Is this estimator consistent?
HINT: USE (9.16) □

9.6 The McDonald–Kreitman Test

The MCDONALD and KREITMAN (1991) test is similar to the HKA test in that it compares the levels of polymorphism and divergence at two sets of sites. Whereas for the HKA test the two sets of sites are two different loci, the McDonald-Kreitman test examines sites that

are interspersed: synonymous and nonsynonymous sites in the same locus. Because the sites are interspersed, it is safe to assume that the genealogies for the two are the same. The model therefore has four parameters; the synonymous mutation rate μ_s and nonsynonymous mutation rate μ_a, the total lengths of divergence and within species branches in the genealogy, t_d and t_w (see Figure 9.11).

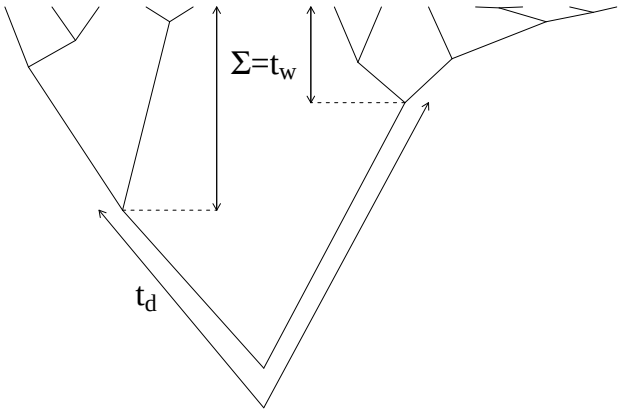


Figure 9.11: Model assumed for the McDonald-Kreitman test.

Assuming the infinite sites model, the numbers of segregating sites of each of the four possible types (synonymous or nonsynonymous, and diverged or polymorphic within a species) are independently Poisson distributed with means given in Figure 9.12.

Here $\mu = \mu_a + \mu_s$ is the total mutation rate and $t = t_d + t_w$ is the total length of the tree. The observations can therefore be arranged in the corresponding 2×2 contingency table and tested for goodness of fit using a χ^2 test, Fisher's exact test or others (e.g. a G-test).

The McDonald-Kreitman test is a very robust test, because no assumption about the

	diverged	polymorphic	Total
synonymous	$\mu_s t_d$	$\mu_s t_w$	$\mu_s t$
non-synonymous	$\mu_a t_d$	$\mu_a t_w$	$\mu_a t$
Total	μt_d	μt_w	μt

Figure 9.12: Expected means for the McDonald-Kreitman test under neutrality.

	diverged	polymorphic	Total
synonymous	17	42	59
non-synonymous	7	2	9
Total	24	44	68

Figure 9.13: Data from McDonald and Kreitman (1991) given in a 2×2 contingency table.

shape of the genealogy is made. It is therefore insensitive to the demographic histories, geographic structuring and non-equilibrium statuses of the populations sampled.

If synonymous mutations are considered neutral on an *a priori* basis, then a significant departure from independence in the test is an indicator of selection at nonsynonymous sites. An excess of substitutions can be interpreted as evidence of adaptive evolution, and an excess of polymorphism can be interpreted as evidence of purifying selection since deleterious alleles contribute to polymorphism but rarely contribute to divergence.

McDonald and Kreitman (1991) studied the *Adh* gene using 12 sequences from *D. melanogaster*, 6 from *D. simulans* and 12 from *D. yakuba* and obtained the data as given in Figure 9.13. These show a significant departure from independence with $p < 0.01$. The departure is in the direction of excess non-synonymous divergence, providing evidence of adaptively driven substitutions.

A R: a short introduction

R is a programming environment for statistical analysis. It is already widely-used in all kinds of statistical analysis, as well as in bioinformatics. Importantly, it is free and works on all major operating systems. The main advantages of R are threefold: First, it is easily extendible, and can be used for both data analysis and as a scripting language. Second, it is relatively easy to learn. Third, it has excellent capabilities for graphical output.

During our course we are using the (seld-written) R-package `labpopgen` for several computations and simulations. Since R is a command line program, we show you how to use it here. R is free software, so you can find many introductions, manuals, etc in the internet; you might e.g. look at `www.r-project.org` which is the homepage of the program.

Once you launched R you find yourself in from of the command prompt

```
>
```

To get started, let us use R as a calculator. So, you can add some numbers and hitting `return` gives you the answer:

```
> 1+2+3+4
[1] 10
```

Within R you can define variables. E.g. you can say

```
> a=1
> b=2
> c=3
> d=4
> a+b+c+d
[1] 10
```

(Often, the sign `<-` is used instead of `=`, e.g. `a<-1`). R takes everything you give it as a vector. In particular, the answer to your last computation was a vector of length 1, which is indicated by the `[1]` in the last line. To define a vector of length greater than one, you can use the `c` command, which stands for 'concetenate'. Moreover, you can compute with these vectors, e.g.

```
> v<-c(1,2,3,4)
> sum(v)
[1] 10
```

Assume you do not know what the function `sum` in the last example really does. Since R comes with a lot of help files, you can just ask the program in typing

```
>?sum
```

and you get a detailed description. Type `q` to exit the help-mode.

`R` is a statistical package, so you can also perform statistical tests. Maybe you know the t-test or the χ^2-test or Fisher's exact test from some statistics course. If you don't know the command for such a test you might find it out, e.g. by saying

```
>help.search("Fisher's exact test")
```

which we will use in Section 9.

In the simulations we will do you will see that `R` can produce nice graphical output. Assume that you measure the weight and height of all persions in class. You might get

```
>height=c(170,174,190,178,163,176,167,165,188,182)
>weight=c(60,67,87,60,51,63,64,59,80,75)
```

You can get a good overview of the data if you see it graphically, so type

```
>plot(weight,height)
```

We wrote some `R`-functions for the course. This package requires another package, `odesolve`, which you should download and install first. (For installing a package, type

```
>R CMD ISNTALL package-name.tgz/zip
```

in a command line under `Linux` or by clicking under `Windows`.) Everytime you launch `R` type

```
>library(labpopgen)
```

whicih then loads all commands which are used in the exercises. You can e.g. find out the details of the coalescent-simulator if you type `?coalator`. To find out all commands in the package type `library(help=labpopgen)`.

Good luck!!!

References

Durrett, R., 2002 *Probability Models for DNA Sequence Evolution*. Springer.

Ewens, W., 2004 *Mathematical Population Genetics*. Springer.

Eyre-Walker, A., N. Smith, and J. Maynard-Smith, 1999 How clonal are human mitochondria? Proc. R. Soc. Lond. B **266:** 477–483.

Fay, J. and C. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Felsenstein, J., 2004 *Inferring Phylogenies*. Sinauer.

Fu, Y.-X. and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gavrilin, G., E. Cherkasova, G. Lipskaya, O. Kew, and V. Agol, 2000 Evolution of Circulating Wild Poliovirus and a Vaccine-Derived Polyivirus in an Immunodeficit Patient: a Unifying Model. Journal of Virology *74*(16)**:** 7381–7390.

Gillespie, J., 2004 *Population Genetics. A Concise Guide* (2nd ed.). John Hopkins University Press.

Haigh, J., 1978 The accumulation of deleterious genes in a population–Muller's Ratchet. Theor. Popul. Biol. *14*(2)**:** 251–267.

Halliburton, R., 2004 *Introduction to Population Genetics*. Prentice Hall.

Hammer, M., D. Garrigan, E. Wood, J. Wilder, Z. Mobasher, A. Bigham, J. Krenz, and M. Nachman, 2004 Heterogeneous Paterns of Variation Among Multiple Human X-Linked Loci: The Possible Role of Diversity-Reducing Selection in Non-Africans. Genetics **167:** 1841–1853.

Hartl, D. and A. Clark, 2007 *Principles of Population Genetics* (4th ed.). Sinauer.

Hedrick, P., 2005 *Genetics of Populations* (3rd ed.). Jones and Bartlett.

Hudson, R., M. Kreitman, and M. Aguade, 1987 A test of neutral evolution based on nucleotide data. Genetics **116:** 153–159.

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics *132*(2)**:** 583–589.

Johnson, T., 2005 Detecting the Effects of Selection on Molecular Variation. Manuscript.

Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Maynard Smith, J. and J. Haigh, 1974 The hitchhiking effect of a favorable gene. Genetic Research **23:** 23–35.

McDonald, J. and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652–654.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press.

Schlenke, T. and D. Begun, 2003 Natural Selection Drives Drosophila Immune System Evolution. Genetics *164*(4)**:** 1471–1480.

Sokal, R. and F. Rohlf, 1994 *Biometry* (3rd ed.). W.H. Freeman.

Tajima, F., 1989 Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics **123:** 585–595.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7:** 256–276.

Index

GNU Free Documentation License
Version 1.2, November 2002

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship

could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the

meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at

least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified,

and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations

of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.