Introduction to Population Genetics

Lecture Notes

Joachim Hermisson

October 12, 2024

University of Vienna
Mathematics Department
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria

Copyright (c) 2024 Joachim Hermisson. For personal use only - not for distribution. These lecture notes are based on previous course material from the "Population Genetics Tutorial" by Peter Pfaffelhuber, Pleuni Pennings, and JH (2009), and from the lecture "Mathematical Population Genetics" by Reinhard Bürger and JH (2015).

Literature

- Sarah P. Otto and Troy Day (2007). A Biologist's Guide to Mathematical Modeling. Princeton University Press, Princeton. *Comprehensive and well-written textbook for students with a biology background.*

- Sean H. Rice (2004). Evolutionary Theory. Sinauer, Sunderland. *Well-written didactic introduction into the field.*

- John Wakeley (2009). Coalescent Theory. Roberts & Co., Greenwood Village. *Comprehensive introduction into the coalescent.*

- Brian Charlesworth and Deborah Charlesworth (2010). Elements of Evolutionary Genetics. Roberts & Co., Greenwood Village.

- Reinhard Bürger (2000). The Mathematical Theory of Selection, Recombination, and Mutation. Wiley, New York. *Advanced textbook for deterministic modelling approaches in population genetics.*

- Warren J. Ewens (2004). Mathematical Population Genetics. Springer, New York. *Advanced textbook for stochsastic modelling approaches in population genetics.*

What is population genetics? – Basic concepts and definitions

Evolution describes the change in heritable characteristics of biological populations over time. Depending on the type of these characteristics, and depending on the time-scale of interest, we can distinguish different branches of evolutionary research.

- *Phylogenetics* is the study of constructing of the *tree of life*, following Darwin's insight that all life on Earth (and the fossil record) shares common ancestors. Changes in traits and characteristics among species, or the emergence of new traits, occur over macroevolutionary timescales, millions or billions of years. Differences between individuals within each species can usually be ignored relative to the differences between species. Each species is therefore usually represented by only a single data point, such as a consensus DNA sequence ("the" human genome).

- *Population genetics* and *quantitative genetics* are interested in the microevolutionary process within a population. Microevolution deals with heritable characteristics that differ among individuals in a population, and describes how the distribution of these characteristics changes across generations. Going back to Darwin (again) and to Wallace, the elementary forces that drive these changes are well-understood: mutation, selection, recombination, genetic drift, and gene flow/migration. Unlike phylogenetics, which is essentially a historical science, microevolution has a mechanistic basis that can be used to construct theoretical models and to make predictions about the future.

Genotype and phenotype

Each individual in a population can be characterized by a large number of morphological, physiological, and behavioral traits, which collectively define its *phenotype*. Individual phenotypes may be more or less adapted to the environmental conditions and influence the viability or reproductive success of their carriers. As a consequence, selection operates on phenotypes. Phenotypes themselves are not inherited, but phenotypic characteristics, such as body size, are influenced by heritable and non-heritable factors.

The part of an individual that is directly heritable is its *genotype*. The genotype of each individual is largely encoded in its *genome* and is represented by its DNA (DeoxyriboNucleic Acid) sequence. DNA is a polymer made up of four types of nucleotides, which differ in the *base* that they contain: adenine (A), guanine (G), thymine (T), and cytosine (C). The nucleotides are organized into two polynucleotide chains that form a double helix with A-T and G-C base pairs. In eukaryotic cells (animals, plants, fungi), the cell nucleus contains several such DNA double strands, called *chromosomes*. In prokaryotic cells (bacteria and archaea), DNA typically forms a single ring (bacterial chromosome). Through development, the genotype determines (the heritable part of) the phenotype, but the connection is extremely complex for most phenotypic traits. The genotype naturally decomposes into *genes*, functional units of DNA that code for a single protein. Quantitative traits of interest (such as milk yield in cows) are usually influenced by a large number of genes.

Due to the complexity of the genotype-phenotype map, all models of (micro)evolution must rely on simplifying assumptions. Quantitative genetic models rely on phenotypic data, but often do not resolve individual genes. Rather, they infer heritable and non-heritable parts of phenotypic traits directly from trait measurements across generations. Population genetic models, on the other hand, directly track the frequencies of genotypes and variants of genes in a population. They often do not refer to phenotypes at all, but assume that selection acts directly on the genes, regardless of where the selection pressure comes from and how it is transmitted across the genotype-phenotype map.

Genes, loci, and alleles

Population genetics is concerned with the evolutionary dynamics of genotypes. It follows the frequencies of genetic variants or *alleles* that differ between individuals. The complete genotype of each individual is given by its DNA sequence ($\approx$ 3 billion base pairs in the human genome, $\approx$ 130 million in *Drosophila*). Usually however, one is only interested in certain aspects of the genotype, such as the genomic positions, or *genetic loci*, that affect a phenotypic trait of interest. At the molecular level, a locus is the position of a single base in the DNA. There are four alleles, corresponding to the four different bases, A, T, G, and C. However, often the term locus is used at a coarser level to refer to the position of a gene or some other significant stretch of sequence. It is always assumed that a locus is a "unit of recombination" that is not broken up during reproduction. There can be many different alleles at a single locus (4^n different alleles for a gene that is represented by a

DNA sequence of fixed length n), but usually one considers classes of equivalent alleles. Many population genetic models distinguish only two classes: an ancestral *wildtype* and a *mutant* allele.

Genetic loci can have different levels of *ploidy*. Most simple life forms (bacteria, mosses, algae, fungi) have a single copy of each chromosome, they are *haploid*. For haploids, a single-locus genotype is determined by a single allele. Almost all higher plants and animals are *diploid*, i.e., most of their chromosomes (the so-called *autosomes* = non-sex chromosomes) are present twice in each adult cell. Some organisms (mostly plants) have an even higher ploidy level (e.g., *tetraploid* with a 4-fold set). Consequently, single-locus genotypes in diploids are given by a pair of alleles (4 alleles in tetraploids, etc).

Mathematical methods

The art of mathematical modeling is to choose the appropriate mathematical methods for the scientific question at hand. Since population genetics is concerned with changes in allele frequencies as a function of time, natural mathematical methods come from fields that describe such processes. Often, the most important decision for a given problem is to decide whether a deterministic or a stochastic framework is appropriate.

- Deterministic models in population genetics use methods from the theory of dynamical systems and of differential equations. On the biological side, this is appropriate if stochastic effects due to a finite population size (genetic drift) can be ignored. This is usually the case if selection is the dominant population genetic force and if the total number of individuals carrying a particular allele is not very small. The dynamics can be modeled in discrete time (using discrete dynamical systems) when a generation is a natural time unit in the biological system, as in annual plants. In other cases, a continuous-time dynamics (based on differential equations) is more appropriate and/or more convenient.

- If genetic drift has a strong effect on the evolutionary process, stochastic models are needed. Basically all these models build on Markov processes (assuming that evolution is only affected by the current state of the population, not its entire history)- Typical examples are birth-death processes or branching processes. As in the deterministic case, they can proceed in discrete or in continuous time. In particular, coalescent theory is a stochastic process which proceeds backwards in time, from the present to the past. This turns out to be particularly useful if we want to explain observed patterns of diversity in data by past evolutionary processes. If population sizes are large and if selection is not too strong, allele frequencies can be treated as a continuous random variable on the unit interval. This leads to diffusion processes as a model of evolutionary change. In fact, parts of the theory of diffusion were developed in the early 20th century with applications in population genetics in mind.

1 Deterministic evolutionary dynamics

1.1 Selection at a single haploid locus

Consider a haploid population of size N. We characterize the genotype by the allelic type at a single locus. There are k alleles, denoted $\{A_1, \ldots, A_k\}$. Generations are discrete and we assume that the population is sufficiently large that stochastic effects due to genetic drift can be ignored. Assume that there are initially n_i individuals with allele A_i. The frequency of A_i in the population is thus $p_i = n_i/N$. Reproduction is clonal, offspring inherit the genotype of their (single) parent, without any modification (no mutation). We are interested in the change of allele frequencies due to selection across a single generation.

Fitness

The fundamental property of individuals that leads to selection and drives adaptive evolution is their fitness. In population genetics, we assign fitness values directly to genotypes or alleles, as follows:

- The *viability* $v_i \geq 0$ measures the probability that a newborn A_i individual survives to reproductive age ($v_i = 0$ means that the individual is inviable).
- The *fecundity* $f_i \geq 0$ measures the expected number of offspring of an adult A_i individual ($f_i = 0$ means that the individual is sterile).
- Finally, the *(absolute) fitness* of allele A_i is defined as

  $$w_i = v_i \cdot f_i \,.$$

  $w_i \geq 0$ measures the expected number of offspring of a newborn A_i individual. Ignoring stochastic effects, we thus have $n_i' = w_i n_i$ for the number n_i' of A_i individuals in the next generation.

For the change in a single generation, we obtain

$$N' = \sum_i n_i' = \sum_i w_i n_i = \Big(\sum_i w_i p_i\Big) N =: \bar{w} N \tag{1.1}$$

where $\bar{w} = \sum_i p_i w_i$ is the *mean fitness* in the population. For the change in allele frequencies, the canonical selection equation for a single haploid locus follows as

$$p_i' = \frac{n_i'}{N'} = \frac{w_i n_i}{\bar{w} N} = \frac{w_i}{\bar{w}}\, p_i \quad \text{or:} \qquad \Delta p_i = p_i' - p_i = \frac{w_i - \bar{w}}{\bar{w}}\, p_i \,. \tag{1.2}$$

We see that any fitness differences among alleles that are represented in the population ($w_i \neq w_j$ for $p_i, p_j > 0$) entails evolutionary change due to selection.

For allele frequency changes across multiple generations, we need to account for the fact that absolute fitness values, as defined above, are usually not constant across generations. Indeed, $w_i = w_i(N)$ is usually not only a function of the allelic type A_i, but (at least) also of the population size N (or density).

- Imagine first that fitness does not depend on the population size (or density). We then have $n_i' = w_i n_i$ and either obtain unlimited growth or decline of n_i over multiple generations, or (with $w_i = 1$ for all alleles) no selection. This is clearly unrealistic.

- Assume next that fitness does depend on density, but not on the allelic state. We then have $\bar{w}(N) = w_i(N) =: w(N)$ and thus

$$p_i' = p_i \quad ; \quad N' = w(N)N\,.$$

  This means we have only changes in the population size (population dynamics), but no changes in the allele frequencies (population genetics) and thus no evolution. Pure population dynamics is a topic of theoretical ecology. With models like logistic growth ($w(N) = r - cN$), population sizes can be regulated and converge to a finite, no-zero value.

- To obtain a reasonable evolutionary model, we need to combine a model of population regulation with a model of evolutionary change. A canonical approach that is implicit to most models in population genetics is to assume that population size regulation is independent of selection. Absolute fitness values then decompose into two parts

$$w_i(N) := w(N) \cdot w_i\,,$$

  (where we distinguish 3 "w functions", $w_i(N)$, $w(N)$, and w_i, in slight abuse of notation). This leads to

$$p_i' = \frac{w_i(N)}{\bar{w}(N)}\,p_i = \frac{w(N)w_i}{w(N)\sum_i w_i p_i}\,p_i = \frac{w_i}{\sum_i w_i p_i}\,p_i = \frac{w_i}{\bar{w}}\,p_i \tag{1.3}$$

  and the density dependence drops out. Following this idea, population genetic models usually do not work with absolute fitness values, but only the *relative fitness* values. If population size regulation is independent of selection, relative fitnesses are density independent. We can then ignore changes in the population size in population genetic models and only follow the dynamics of allele frequencies. Note that, from here on, we will use the symbols w_i and $\bar{w}$ for (mean) relative fitness only. Likewise, if we simply refer to *fitness*, it is relative fitness what is meant.

- Since any factor that is common to all fitness values w_i drops out of the selection equation, relative fitness values w_i are only defined up to a constant factor. We can use this freedom to normalize the fitness of some reference allele A_1 (often: the ancestral wildtype allele) to $w_1 = 1$.

- Following these leads, the easiest model of selection results if we assume constant *relative* fitness values for all alleles. The change in p_i across t generations follows as

$$p_i(t) = \frac{n_i(t)}{N} = \frac{w_i^t\, n_i(0)}{\sum_j w_j^t\, n_j(0)} = \frac{w_i^t\, p_i(0)}{\sum_j w_j^t\, p_j(0)}\,. \tag{1.4}$$

If $w_1 > w_j,\ j \geq 2$, we obtain

$$p_i(t) = \frac{p_i(0)}{\sum_j (w_j/w_i)^t \cdot p_j(0)} \quad \xrightarrow{t \to \infty} \quad \frac{p_i(0)}{p_1(0) \lim_{t\to\infty} (w_1/w_i)^t} = \delta_{1,i}\,.$$

We conclude that with constant (time-homogeneous and frequency-independent) selection in haploids only the fittest allele survives and fixes in the population. There is no genetic variation maintained.

1.2 Selection at a single diploid locus

Consider a diploid locus with two alleles (wildtype and mutant), A and a. In principle, there can be $2 \times 2 = 4$ genotypes at the locus, but if there is no *position effect* (i.e. it does not matter on which DNA strand an allele is located), there are only three: the two *homozygous* genotypes AA and aa and the *heterozygous* genotype Aa $(= aA)$. Let x, y, and z be the frequencies of genotypes AA, Aa, and aa, respectively. We can express the frequencies $p = x + y/2$ of the A allele and $q = z + y/2$ of the a allele in terms of the genotype frequencies, but note that this is generally not possible vice-versa.

Random mating and Hardy-Weinberg proportions

To describe evolutionary dynamics in diploids, even without selection, we first need a model for the change in genotype frequencies under reproduction. Most diploids reproduce sexually. Under *Mendelian inheritance*, each newborn inherits a single allele from both parents at each autosomal locus. In general, the change of genotype frequencies across generations depends on the mating pattern. For example, males and females often prefer mating partners with similar phenotypic characteristics such as body size (assortative mating). However, the simplest mating scheme that is also used *by default* in population genetics assumes that matings are random. We also assume that sexes are equivalent and there are no differences in genotype frequencies among males and females in the population (this is necessarily true for monoecius species, where all individuals act in male and female roles). We can then summarize the offspring frequencies for each mating type in a table:

♀	♂	mating prob.	x'	y'	z'
AA	AA	x^2	1	0	0
	Aa	xy	1/2	1/2	0
	aa	xz	0	1	0
Aa	AA	xy	1/2	1/2	0
	Aa	y^2	1/4	1/2	1/4
	aa	yz	0	1/2	1/2
aa	AA	xz	0	1	0
	Aa	yz	0	1/2	1/2
	aa	z^2	0	0	1

$$x' = 1 \cdot x^2 + 2\,\frac{1}{2}\,xy + \frac{1}{4}\,y^2 = \Big(x + \frac{y}{2}\Big)^2 = p^2$$

$$y' = 2\,\frac{1}{2}\,xy + 2\,\frac{1}{2}\,yz + 2xz + \frac{1}{2}\,y^2$$

$$= 2\Big(x + \frac{y}{2}\Big)\Big(z + \frac{y}{2}\Big) = 2pq$$

$$z' = 1 \cdot z^2 + 2\,\frac{1}{2}\,yz + \frac{1}{4}\,y^2 = \Big(z + \frac{y}{2}\Big)^2 = q^2$$

The third column of the table gives the probability of the mating pair under random mating and columns 4 to 6 the genotype frequencies in the offspring generation under Mendelian inheritance, conditioned on the mating pair. The total (unconditioned) genotype frequencies in the offspring generation derived by summing over all mating pairs. We observe:

- The genotype frequencies after a single generation of random mating are determined by the allele frequencies, $(x', y', z') = (p^2, 2pq, q^2)$: *Hardy-Weinberg proportions.*

- The allele frequencies do not change under random mating

$$p' = x' + \frac{1}{2}y' = p \quad ; \quad q' = z' + \frac{1}{2}y' = q\,.$$

  There is thus no loss of genetic variation under Mendelian inheritance.

- The so-called *Hardy-Weinberg law* states that, after a single generation of random mating, both the allele frequencies and the genotype frequencies remain invariant: They are in *Hardy-Weinberg equilibrium.*

- It is easy to extend the Hardy-Weinberg law to an arbitrary number of alleles $\{A_1, \ldots, A_k\}$. Let $P_{ij} = P_{ji}$ denote the frequency of the genotype A_iA_j. The allele frequency of A_i is $p_i = P_{ii} + \frac{1}{2}\sum_{j\neq i} P_{ij}$. A straight-forward extension of the 2-allele derivation shows that $p'_i = p_i$, $P'_{ii} = p_i^2$ and, for $j \neq i$, $P'_{ij} = 2p_ip_j$.

The important consequence of the Hardy-Weinberg (HW) law for population genetic models is that it is sufficient to follow k allele frequencies, rather than the $k(k+1)/2$ frequencies of diploid genotypes. However, the law is only valid under a number of assumptions.

- Random mating: with other mating schemes (e.g. assortative mating or selfing), we obtain different equilibrium frequencies *and* generally only gradual (asymptotic) convergence to this equilibrium, rather than convergence in a single generation.

- Discrete Generations: Convergence to HW proportions is only asymptotic if generations are overlapping (individuals do not all reproduce and die at the same time).

- Equivalent sexes: If the initial allele frequencies in males and females differ, HW proportions are only reached in two generations of random mating.

- Autosomal loci: For X-linked loci (that are diploid in females, but haploid in males), HW proportions are only reached asymptotically.

- No selection, mutation, or drift: all evolutionary forces readily lead to deviations from HW proportions. However, as we will see below, we can often still make use of the HW law at certain stages of a diploid *life cycle.*

Viability selection at a single diploid locus

Consider a diploid population with discrete generations and equivalent sexes and a single locus with two alleles, A and a with frequencies p and q, respectively. We also assume that selection acts on the viability, the probability that newborn diploid individuals reach reproductive age. We can then dissect the life-cycle of the population into two phases: a selection phase, during which juveniles grow up and a reproductive phase where adults mate and produce offspring. The key assumption is that selection and reproduction can be separated and occur at different stages.

- Consider the reproductive phase first. If reproduction works via random mating as described above, we can use the results of the HW law: Allele frequencies are conserved during the reproductive step and genotype frequencies will be in HW equilibrium directly after reproduction (for zygotes (= newly fertilized eukaryotic cell) not yet affected by selection).

- We still need a model for the change of allele and genotype frequencies during the reproductive phase. We assign fitness values w_{AA}, w_{Aa}, and w_{aa} to the three genotypes AA, Aa, and aa, respectively. The genotypes frequencies are P_{AA}, P_{Aa}, and P_{aa}, and the allele frequencies are $p = P_{AA} + P_{Aa}/2$ and $q = P_{aa} + P_{Aa}/2$. We can the define *marginal fitness* values for the alleles A and a,

$$w_A = \frac{w_{AA}2P_{AA} + w_{Aa}P_{Aa}}{2P_{AA} + P_{Aa}} \quad , \quad w_a = \frac{w_{aa}2P_{aa} + w_{Aa}P_{Aa}}{2P_{aa} + P_{Aa}} \, .$$

  The mean fitness in the population follows as

$$\bar{w} = w_{AA}P_{AA} + w_{Aa}P_{Aa} + w_{aa}P_{aa} = w_A p + w_a q \, .$$

  With these definitions, the changes in genotype and allele frequencies over a life cycle can easily be expressed. They are summarized in the following table.

	AA	Aa	aa	A	a
frequency after random mating	$P_{AA} = p^2$	$P_{Aa} = 2pq$	$P_{aa} = q^2$	p	q
frequency after selection	$p^2 \frac{w_{AA}}{\bar{w}}$	$2pq \frac{w_{Aa}}{\bar{w}}$	$q^2 \frac{w_{aa}}{\bar{w}}$	$p \frac{w_A}{\bar{w}} = p'$	$q \frac{w_a}{\bar{w}} = q'$
next gener. frequency after random mating	$P'_{AA} = p'^2$	$P'_{Aa} = 2p'q'$	$P'_{aa} = q'^2$	p'	q'

  Note that the diploid selection equation for the allele frequencies takes the same functional form as in the haploid case if we replace the allelic fitness value by the corresponding marginal fitness. In general, marginal fitnesses and the mean fitness depend on the genotype frequencies and the equations on the level of allele frequencies do not form a closed dynamical system. In the special case of random mating and viability selection, however, we can express genotype frequencies as HW proportions and the dynamical system for the allele frequencies closes. In particular, the marginal fitness values simplify to

$$w_A = w_{AA}p + w_{Aa}q \quad , \quad w_a = w_{aa}q + w_{Aa}p \, .$$

Selection scenarios

We have seen that viability selection on a single diploid locus with random mating leads to a selection equation that is formally equivalent to the haploid case. The difference is that the marginal fitness values for the alleles depend on the allele frequencies, even if the genotypic fitness values are constant. This leads to differences in the evolutionary dynamics. To characterize these differences, we use the following classical parametrization of the genotypic fitness values.

$$w_{aa} = 1 \qquad \text{normalization of the (relative) wildtype fitness} \tag{1.5a}$$

$$w_{AA} = 1 + s \qquad s\text{: selection coefficient for the homozygote mutant} \tag{1.5b}$$

$$w_{Aa} = 1 + hs \qquad h\text{: dominance coefficient for heterozygote fitness} \tag{1.5c}$$

Depending on the value of the dominance coefficient, we distinguish the following biological scenarios for the mutant allele A

$$h \begin{cases} > 1 & \text{overdominant} \\ = 1 & \text{(fully) dominant} \\ \in (\frac{1}{2}, 1) & \text{partially dominant} \\ = \frac{1}{2} & \text{codominant (or no dominance)} \\ \in (0, \frac{1}{2}) & \text{partially recessive} \\ = 0 & \text{(fully) recessive} \\ < 0 & \text{underdominant} \end{cases}$$

For all cases, the marginal allele fitnesses and mean fitness in HW equilibrium follow as

$$w_a = 1 + p \cdot hs \tag{1.6a}$$

$$w_A = 1 + q \cdot hs + p \cdot s \tag{1.6b}$$

$$\bar{w} = 1 + 2pq \cdot hs + p^2 \cdot s \tag{1.6c}$$

and the allele frequency change per generation of the mutant allele is

$$\Delta p = p' - p = \frac{w_A - \bar{w}}{\bar{w}}\, p = pq\, \frac{s(h + (1-2h)p)}{\bar{w}}\,.$$

In contrast to the haploid case, there is usually no explicit solution for the allele frequency $p(t)$ as a function of time. However, it is straightforward to derive the equilibrium frequencies of the dynamical system. We have $\Delta p = 0$ for

$$p = 0 \quad , \quad p = 1 \;\; [\Leftrightarrow q = 0] \qquad\qquad (\textit{monomorphic equilibria})$$

$$h + (1-2h)p = 0 \quad \Rightarrow \quad p = \hat{p} = \frac{h}{2h-1} \qquad (\textit{polymorphic equilibrium})$$

The equilibrium at $\hat{p}$ is in the interior of the frequency space, $0 < \hat{p} < 1$, if and only if either $h > 1$ (A is overdominant) or $h < 0$ (A is underdominant). We can distinguish three parameter ranges, based on the dominance coefficient, that lead to qualitatively different dynamical behavior.

1. In the whole parameter range $0 \leq h \leq 1$, ranging from complete recessiveness to complete dominance of the mutant allele A, we have

$$h + (1 - 2h)p > 0 \quad \text{for} \quad 0 < p < 1$$

and thus $\Delta p > 0$ for a beneficial mutant ($s > 0$), resp. $\Delta p < 0$ for a deleterious mutant ($s < 0$). The dynamical system therefore converges monotonically either to the equilibrium at $p = 1$ or to $p = 0$ for the beneficial or deleterious case, respectively.

2. If an equilibrium $\hat{p}$ at an intermediate frequency exists, we can write

$$\Delta p = \frac{pqs(2h-1)}{\bar{w}} \, (\hat{p} - p) \, .$$

With $s > -1$, we also have

$$\bar{w} - pqs(2h-1) = 1 + 2pqhs + p^2 s - pqs(2h-1) = 1 + ps > 0 \, .$$

We therefore obtain monotone convergence of $p(t)$ toward the polymorphic equilibrium $\hat{p}$ for the overdominant case ($h > 1$) if $s > 0$ and for the underdominant case ($h < 0$) if $s < 0$. In both cases, the heterozygote is the fittest genotype (*heterozygote advantage*). Note that an underdominant allele A corresponds to an overdominant allele a. The term *overdominance* is often used as synonymous to *heterozygote advantage*, implicitly using the allele with the higher fitness as reference.

3. Analogously, we find monotonic divergence from $\hat{p}$ toward either $p = 0$ or $p = 1$ for the underdominant beneficial case ($h < 0$ and $s > 0$) and for the overdominant deleterious case ($h > 1$ and $s < 0$).

We see that heterozygote advantage ("overdominance") is necessary and sufficient for the maintenance of genetic variation under selection at a single diploid locus.

Multiple alleles

It is easy to extend the 2-alleles case for a single diploid locus to the general case of k alleles, $\{A_1, \ldots, A_k\}$ with frequencies $\{p_1, \ldots, p_k\}$. Let $w_{ij} = w_{ji}$ be the fitness value of genotype $A_i A_j$, with frequency P_{ij} in the population. After random mating, the population is in HW equilibrium, thus $P_{ii} = p_1^2$ and $P_{ij} = 2p_i p_j$ for $i \neq j$. The marginal allelic fitnesses and the mean fitness are

$$w_i = \sum_j w_{ij} p_j \quad , \quad \bar{w} = \sum_i w_i p_i = \sum_{i,j} w_{ij} p_i p_j$$

and the change in allele frequencies is

$$p_i' = \frac{w_i}{\bar{w}} \, p_i \qquad \text{resp.} \qquad \Delta p_i = p_i' - p_i = \frac{w_i - \bar{w}}{\bar{w}} \, p_i \, . \tag{1.7}$$

Continuous time model for selection

Mathematically, our model so far for the evolutionary dynamics has been a discrete dynamical system. A model in discrete time is realistic for some biological species (e.g. annual plants), it has some technical advantages (in particular, it allows for a separation of reproduction and selection) and it is easy to simulate on a computer. However, it is often more convenient (and/or more realistic biologically) to model evolution in continuous time. To this end, consider again a single haploid locus with k alleles, $\{A_1, \ldots, A_k\}$. If births and deaths occur at a constant rate, the number n_i of A_i types changes like

$$\dot{n}_i(t) = \frac{dn_i(t)}{dt} = (b_i - d_i)n_i(t) = m_i n_i(t), \tag{1.8}$$

where b_i and d_i are the birth- and death-rates and $m_i = b_i - d_i$ the total growth rate, which is also called the *Malthusian fitness* of allele type A_i. The dynamics of the total population follows as

$$\dot{N}(t) = \sum_i \dot{n}_i(t) = \sum_i m_i n_i(t) = N(t) \sum_i m_i p_i(t) = \bar{m}(t)N(t), \tag{1.9}$$

where $p_i(t) = n_i(t)/N(t)$ is the frequency of allele A_i and $\bar{m}(t) = \sum_i m_i p_i(t)$ the mean Malthusian fitness. The allele frequencies change according to

$$\dot{p}_i(t) = \frac{d}{dt}\left(\frac{n_i(t)}{N(t)}\right) = \frac{N(t)\dot{n}_i(t) - n_i(t)\dot{N}(t)}{N^2(t)} = \big(m_i - \bar{m}(t)\big)p_i(t). \tag{1.10}$$

Since Eq. (1.8) implies exponential growth (or decline) of the n_i, the Malthusian fitness values m_i (like the "Wrightian" fitness values w_i in discrete time) need to depend on the population density in a realistic model. However, as in discrete time, this dependence drops out for the allele frequency dynamics Eq. (1.10) if we assume the same density dependence for all alleles (and set $m_i(N) = m(N) + m_i$).

- If we assume Hardy-Weinberg proportions, the haploid evolution equation can again be generalized to diploids with fitness values m_{ij} for genotype A_iA_j, marginal fitness $m_i = \sum_j m_{ij}p_j$ and mean fitness $\bar{m} = \sum_{ij} m_{ij}p_ip_j$.

- Whereas the evolution equations in discrete time take the form of *difference equations*, they are *ordinary differential equations* (ODEs) in continuous time. Since ODEs are convenient from a mathematical perspective, they are often preferred in models. For diploids, however, the formalism is only approximate. This is because selection in continuous time causes deviations from HW equilibrium (unless Malthusian fitness is additive, $m_{ij} = m_i + m_j$, corresponding to the assumption of no dominance). Since deviations are small for weak selection, both formalisms usually produce equivalent results.

Sir Ronald A. Fisher, 1890–1962, is well-known for both his work in statistics and genetics. He is one of the founding fathers of population genetics (together with JBS Haldane and S Wright) that combined Darwinian selection and Mendelian inheritance in the so-called *Modern Synthesis* and led to the breakthrough of Darwinism in the early 20th century. Fisher's 1930 article on *The Genetical Theory of Natural Selection* defined large parts of the field.
In statistics, Fisher's key achievement was his invention of the analysis of variance, or ANOVA. This statistical procedure allows to connect the observed deviations in experimental data to different controlled and uncontrolled underlying factors. It constituted a notable advance over the prevailing procedure of varying only one factor at a time in an experiment. Fisher summed up his statistical work in his book Statistical Methods and Scientific Inference (1956). Fisher became Galton Professor of Eugenics at University College, London in 1933. From 1943 to 1957 he was Balfour Professor of Genetics at Cambridge. He was knighted in 1952 and spent the last years of his life conducting research in Australia (adapted from Encyclopedia Britannica and Wikipedia).

- We can show that mean Malthusian fitness is non-decreasing. For diploids,

$$\begin{aligned}\dot{\bar{m}} &= \sum_{ij} m_{ij}(\dot{p}_i p_j + p_i \dot{p}_j) = 2\sum_{ij} m_{ij} p_i p_j (m_i - \bar{m}) \\ &= 2\sum_i p_i m_i (m_i - \bar{m}) = 2\sum_i p_i (m_i - \bar{m})(m_i - \bar{m}) = 2V_G \end{aligned} \tag{1.11}$$

  where $V_G > 0$ is the *genetic variance in fitness*. We thus see that the increase in mean fitness is given by (twice) the current variance in fitness in the population. This is the assertion of *Fisher's fundamental theorem of natural selection* that goes back to R.A. Fisher (1930) and has been discussed in many population genetic textbooks. However, this theorem is only exact for a single locus in continuous time and only holds approximately for discrete time and in more general evolutionary situations.

1.3 Mutation-selection models

The ultimate source of all genetic variation in a population is mutation. So far, we have just assumed that genetic variation exists and have not modeled its creation explicitly. Since selection is usually a much stronger force that mutation and leads to allele frequency changes over shorter time scales, this is often a reasonable approximation. However, for a more complete description of evolution over longer time scales, we need to include mutation into the model.

Only mutation

Usually, mutation occurs during reproduction (or: the production of gametes) due to errors in DNA copying. Each generation, there is a probability that an offspring individual does

not inherit the allelic state of (one of) its parent(s), but rather a mutated allele. For a single locus and two alleles, A and a, assume that there is a fixed probability μ that an ancestor carrying the ancestral allele a produces an offspring with A allele. Vice-versa, there is a probability ν that A mutates back to a during reproduction. If the frequency of A alleles is p, the single-generation dynamics reads

$$\Delta p = p' - p = \mu(1-p) - \nu p \tag{1.12}$$

with equilibrium ($\Delta p = 0$)

$$p = \hat{p} = \frac{\mu}{\mu + \nu}\,.$$

For an arbitrary number of alleles $A_1, \ldots, A_k$ and mutation probability from allele A_i to allele A_j denoted as μ_{ij}, the mutation equation reads

$$p_i' = \Big(1 - \sum_j \mu_{ij}\Big) p_i + \sum_j \mu_{ji} p_j\,. \tag{1.13}$$

The analogous equation in continuous time is

$$\dot{p}_i = \sum_j \big(u_{ji} p_j - u_{ij} p_i\big), \tag{1.14}$$

where u_{ij} are mutation rates per unit time.

Combining mutation and selection

In discrete time, we can simply include mutation as a separate step into the life cycle. We define the allele frequency change during one generation, starting with newborn zygotes, as $p_i \to p_i^{(s)} \to p_i'$ with

$$p_i' = \Big(1 - \sum_j \mu_{ij}\Big) p_i^{(s)} + \sum_j \mu_{ji} p_j^{(s)} \quad ; \quad p_i^{(s)} = \frac{w_i}{\bar{w}}\, p_i\,. \tag{1.15}$$

The scheme applies to both haploids and (random mating) diploids, with w_i as marginal fitness for diploids. The first step accounts for viability selection, the second step for mutation during reproduction. It is easy to check that mutation in HW equilibrium changes the allele frequencies, but maintains HW proportions.

There are various ways to write down a mutation-selection equation in continuous time. The most widely used formalism is simply to assume that mutation and selection are independent processes that occur in parallel. This leads to the differential equation

$$\dot{p}_i = (m_i - \bar{m}) p_i + \sum_j \big(u_{ji} p_j - u_{ij} p_i\big) \tag{1.16}$$

combining Eqs. (1.10) and (1.14). The m_i are Malthusian fitness values (marginal fitnesses for diploids) and the μ_{ij} have the interpretation of mutation rates per time unit.

Mutation-selection balance

We can now ask how the combined action of mutation and selection changes the evolutionary dynamics that we have classified above for selection on a single diploid locus. In particular, we are again interested in the equilibrium points that will be reached. Although the general mutation-selection equation allows for quite complex dynamics, things are easier in the biologically most relevant regime, where mutation is much weaker than selection. In this case, mutation acts as a perturbation of the selection dynamics and (only) leads to slight shifts of the equilibrium points. This is still relevant, however, for all cases where selection alone leads to a monomorphic equlibrium that is turned into a stable polymorphism by recurrent mutation.

For simplicity, we consider a single haploid locus in continuous time, with wildtype allele a and deleterious mutant A, with fitnesses 1 and $1-s$, respectively. The mutant A is generated by recurrent mutation at rate u. Since mutants are rare and beneficial mutation from mutant to wildtype is an unlikely event, we can ignore back mutation from A to a. The dynamical equation for the mutant allele frequency p then reads

$$\dot{p} = (m-\bar{m})p + u(1-p) = \big(1-s-(1-ps)\big)p + u(1-p) = -sp(1-p) + u(1-p)\,. \quad (1.17)$$

Setting $\dot{p}=0$, we obtain the solutions $p=1$ and the non-trivial stable equilibrium at

$$\hat{p} = \frac{u}{s}\,. \quad (1.18)$$

Analogous results apply in discrete time and for diploids (with the heterozygous fitness hs replacing s). The equilibrium mean fitness in the population follows as

$$\hat{\bar{m}} = 1 - \hat{p}s = 1 - u\,. \quad (1.19)$$

The difference between the mean fitness and the maximal fitness in the population is also called the mutation load L_m. We thus have

$$L_m = 1 - (1-u) = u\,. \quad (1.20)$$

Whereas the selection coefficient s (or by hs in a diploid heterozygote) serves as a measure for the effect of a deleterious mutation on an individual, the mutation load L_m can be seen as a measure of the effect of deleterious mutation on the population level. We see that the mutation load depends (to leading order) only on the mutation rate, but not on the fitness effects of the deleterious mutations. The reason is that a milder mutation with small s will segregate at a higher frequency $\hat{p}=u/s$ in the population. To leading order, the effects of mutation frequency and mutation size on the mean fitness just cancel. This is also called *Haldane's rule* or the *Haldane-Muller principle* and has relevant consequences for programs of public health that aim for an increase of population-level parameters like the mean fitness. Indeed, according to the Haldane-Muller principle, the mean fitness in a population is neither altered by eugenics (birth control for diseased people, effectively increasing the deleterious fitness effect of a mutation) nor by a partial cure of a genetic disease (reduction of s). For population-level fitness, mildly deleterious mutations are as harmful as strongly deleterious ones. Only the reduction of mutation *rates* has a lasting effect on mean fitness.

JBS (John Burdon Sanderson) Haldane, 1892–1964, was a British geneticist, biometrician, physiologist, and popularizer of science who opened new paths of research in population genetics and evolution. Together with R.A. Fisher and Sewall Wright, but in separate mathematical arguments, he 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.
Haldane 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 (adapted from Encyclopedia Britannica).

Herrmann Joseph Muller 1890–1967, Nobel laureate in Medicine (1946) for his discovery of the mutagenic effect of X-rays was very concerned about the reduction of mean fitness in humans by radiation, also due to nuclear fallout caused by nuclear testing. Together with fellow scientists, he was a vocal critic of nuclear weapons testing (from Wikipedia).

1.4 Recombination

In diploid organisms recombination happens during *meiosis* (the production of gametes). Recombination mixes paternal and maternal material before it is transferred 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, consider your two chromosomes number 1, one of which came from your father and one from your mother. The one that stems from your father is in fact a mosaic of pieces from his mother and his father, i.e., your two paternal grandparents. In humans, these mosaics chromosomes typically consist of 2-10 chunks or recombination blocks. Chromosomes that do not recombine are not mosaics. The Y-chromosome does not recombine at all, males inherit it completely from their father and paternal grandfather, etc. Mitochondrial DNA also does not normally recombine, both females and males inherit mitochondria from their mother, maternal grandmother, etc. The X-chromosome only recombines when it is in a female.

There are various mechanisms for recombination. The most well-known one is *crossing over*, where matching regions in homologous chromosomes (which pair during meiosis) experience a *double strand break* and subsequently are reconnected to the other chromosome. There are other recombination mechanisms like *gene conversion*, where a stretch of DNA is copied from one chromosome to the matching region of its homologous partner. Exchange of genetic material can also happen in haploid individuals. In this case two different individuals exchange pieces of their genome.

Linkage

Mendel's second law (of *independent assortment*) states that genes are inherited independently of each other. It means that the probability of inheriting a gene at some locus $\mathcal{A}$ from one grandmother is independent of whether or not a gene at a different locus $\mathcal{B}$ has been inherited from the same grandmother. This "law" is generally only true for gene loci that are located on different chromosomes: they are *unlinked*. On the other hand, if genes are on the same chromosome, they are said to be *physically linked*. Linked genes are not inherited independently of each other. In particular, if gene loci are very close to each other, recombination between them is rare and they are typically inherited together. Mathematically, this is expressed by the *recombination fraction* $r = r_{AB}$ between loci $\mathcal{A}$ and $\mathcal{B}$, which defines the probability that genes inherited from different grandparents at these loci end up on the same parental gamete (sperm, egg, pollen) that contributes to the offspring genotype,

$$\binom{a_1 b_1}{a_2 b_2} \quad \longrightarrow \quad \begin{cases} \left.\begin{matrix} a_1 b_1 \\ a_2 b_2 \end{matrix}\right\} & \text{freq. } \tfrac{1}{2}(1-r) \text{ each} \\ \left.\begin{matrix} a_1 b_2 \\ a_2 b_1 \end{matrix}\right\} & \text{freq. } \tfrac{1}{2} \cdot r \text{ each} \end{cases} . \tag{1.21}$$

Here, a_1 and b_1 (res. a_2 and b_2) do not denote an allelic state, but only the origin of the gene either from grandparent 1 or 2.

- r is often also called a *recombination rate*, but it is really a probability in discrete generation models. We generally have $r = 1/2$ as upper limit for unlinked loci on different chromosomes and $0 \leq r < 1/2$ for linked loci.

- We can define a molecular recombination probability ρ as the probability for recombination between neighboring base pairs along a chromosome. Typical values are $\rho \approx 10^{-8}$ per generation. However, ρ generally depends strongly on the genomic position x. The estimation of recombination maps $\rho(x)$ from data is an important task of genomics.

- For a given recombination map, we can define a *recombination distance* d along a chromosome in units of *Morgans* (named after Thomas Morgan). A distance of $d = 1M$ indicates that there is on average one recombination breakpoint per generation within the stretch (e.g., due to crossing over). Typical lengths of chromosome regions measure in *centi-Morgans* (cM).

- The recombination fraction r between loci on the same chromosome is the probability of an odd number of recombination breakpoints between these loci. Ignoring interference of recombination events in neighboring regions, r relates to the recombination distance d via *Haldane's mapping function*

$$r = \frac{1}{2}\big(1 - \exp[-2d]\big) . \tag{1.22}$$

Linkage disequilibrium

For simplicity, we focus on the case of two loci, $\mathcal{A}$ and $\mathcal{B}$, with two alleles each, $\{a, A\}$ and $\{b, B\}$. There are then 4 gametes (or haplotypes) $\{ab, aB, Ab, AB\}$, with frequencies denoted $\{P_{ab}, P_{aB}, P_{Ab}, P_{AB}, \}$. The allele frequencies derive as

$$P_a = P_{ab} + P_{aB}\,; \quad P_A = 1 - P_a = P_{Ab} + P_{AB}\,, \tag{1.23}$$

and analogously for the $\mathcal{B}$ locus. As a measure of non-random association of alleles at different loci on the same gamete (or haplotype), we define the *linkage disequilibrium* (LD). E.g., for alleles A and B,

$$\begin{aligned} D_{AB} &= P_{AB} - P_A P_B \\ &= P_{AB}(P_{AB} + P_{Ab} + P_{aB} + P_{ab}) - (P_{Ab} + P_{AB})(P_{aB} + P_{AB}) \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB}\,. \end{aligned} \tag{1.24}$$

It is easy to verify that

$$D := D_{AB} = D_{ab} = -D_{Ab} = -D_{aB}\,,$$

such that LD between two biallelic loci is measured by a single scalar number (this is more complex for more alleles or more loci, see e.g. chapter 5 of the book by R. Bürger). If the linkage disequilibrium is zero, $D = 0$, we say that the alleles are in *linkage equilibrium* (LE). Note that *linkage* and *linkage disequilibrium* are concepts on different levels. While linkage is a property of loci and manifests in each individual, linkage disequilibrium is a population property and related to allele/haplotype frequencies. Unlinked loci can certainly have non-zero linkage disequilibria among their alleles, while alleles at linked loci (even with $r = 0$) can be in linkage equilibrium.

Recombination dynamics

Consider the two-locus model as described above. Without mutation or selection (or drift), the single-locus allele frequencies in the population stay constant, $P'_A = P_A$, etc. However, recombination will change the haplotype frequencies. Assuming HW proportions in the germ cells prior to meiosis (and recombination), we obtain

$$P'_{AB} = (1 - r)P_{AB} + r \cdot P_A P_B = P_{AB} - r \cdot D\,. \tag{1.25}$$

Indeed, a fraction of $(1 - r)$ of all gametes that contribute to the new generation has not undergone any recombination. In this part of the population, haplotype frequencies maintain their value from the previous generation. Conversely, a fraction of r of new gametes are recombination products. In HW equilibrium, the probability for them to result in a AB haplotype is $P_A P_B$. For the change in linkage disequilibrium, we obtain

$$D' = P'_{AB} - P'_A P'_B = (1 - r)P_{AB} + r \cdot P_A P_B - P_A P_B = (1 - r) \cdot D\,. \tag{1.26}$$

- We thus see that for $r > 0$ linkage disequilibrium decays to zero at geometric rate $(1 - r)$. The population approaches linkage equilibrium, $D = 0$, among all alleles.
- Note that, in contrast to HW equilibrium, linkage equilibrium among alleles at different loci is *not* reached in a single generation, but only asymptotically – even for unlinked loci with $r = 1/2$.

Recombination and selection

Consider the following fitness scheme for diploid individuals

$$\begin{array}{c|ccc} & BB & Bb & bb \\ AA & w_{ABAB} & w_{ABAb} & w_{AbAb} \\ Aa & w_{ABaB} & w_{ABab} & w_{Abab} \\ aa & w_{aBaB} & w_{aBab} & w_{abab} \end{array} ,$$

where we assume that the fitness of a genotype depends only on the number and type of alleles in the genotype, but not on the association of the allele to a particular haplotype (no *position effect*). I.e., the fitness of the diploid genotype (Ab, aB) is the same as the one of (AB, ab). Assuming HW proportions in zygote state, marginal fitness values for the 2-locus haplotypes follow in the usual way, $w_{AB} = w_{ABAB}\, P_{AB} + w_{ABAb}\, P_{Ab} + w_{ABaB}\, P_{aB} + w_{ABab}\, P_{ab}$, etc. The mean fitness is

$$\bar{w} = w_{AB}\, P_{AB} + w_{Ab}\, P_{Ab} + w_{aB}\, P_{aB} + w_{ab}\, P_{ab}\,.$$

Like for the mutation-selection model, we can construct a recombination-selection model by including both events as separate steps into a life cycle. Since random mating decomposes whole genotype frequencies into haplotype frequencies, this can be done on the level of haplotypes. Starting with zygotes, we first have selection, followed by recombination during reproduction. This results in

$$P'_{AB} = \hat{P}_{AB} - r\hat{D}\,, \tag{1.27a}$$

$$\begin{aligned} D' &= (\hat{P}_{AB} - r\hat{D})(\hat{P}_{ab} - r\hat{D}) - (\hat{P}_{Ab} + r\hat{D})(\hat{P}_{aB} + r\hat{D}) \\ &= \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} - r\hat{D}\,, \end{aligned} \tag{1.27b}$$

and similar expressions for the other haplotype frequencies. $\hat{P}_{\cdot\cdot}$ and $\hat{D}$ are the values for the frequencies and for LD after selection. We have

$$\hat{P}_{AB} = \frac{w_{AB}}{\bar{w}}\, P_{AB}\,.$$

For $\hat{D}$, we need to consider that recombination occurs in the diploid phase after selection and get

$$\hat{D} = \widehat{P_{AB}P_{ab}} - \widehat{P_{aB}P_{Ab}} = \frac{w_{ABab}}{\bar{w}}\big(P_{AB}P_{ab} - P_{Ab}P_{aB}\big) = \frac{w_{ABab}}{\bar{w}}\, D$$

resulting in

$$D' = \frac{w_{AB}w_{ab}}{\bar{w}^2}P_{AB}P_{ab} - \frac{w_{aB}w_{Ab}}{\bar{w}^2}P_{Ab}P_{aB} - r\frac{w_{ABab}}{\bar{w}}\,D \tag{1.28}$$

Assume that, initially, $D = P_{AB}P_{ab} - P_{Ab}P_{aB} = 0$. Eq. (1.28) shows that selection will create positive or negative LD, depending on the fitness values for haplotypes and on the so-called level of *epistasis*. We have

$$w_{AB}w_{ab} - w_{Ab}w_{aB} \begin{cases} > 0 & \text{positive epistasis, creates positive LD} \quad D' > 0 \\ = 0 & \text{no epistasis, maintains LE} \quad D' = D = 0 \\ < 0 & \text{negative epistasis, creates negative LD} \quad D' < 0\,. \end{cases} \tag{1.29}$$

Epistasis vanishes if all genotype fitnesses are multiplicative across loci ($w_{ABAb} = v_{AA}v_{Bb}$, etc). In this case, the dynamics for $D = 0$ (on the LE manifold) simplifies to an independent single-locus dynamics at both loci,

$$P'_A = (v_A/\bar{v}_A)P_A,$$

with $v_A = v_{AA}P_A + v_{Aa}P_a$ and $\bar{v}_A = v_{AA}P_A^2 + 2v_{Aa}P_AP_a + v_{aa}P_a^2$. The result shows under which conditions the use of simple single locus models is meaningful in complex biological scenarios: if fitness epistasis can be ignored and if loci are in LE. Furthermore, even if starting conditions are not in LE, but $D = 0$ for all equilibria, we can use the single-locus formalism to describe the equilibrium structure and the long-term dynamics.

2 Genetic Drift

In the first part of the lecture, we have described the evolutionary dynamics using a *deterministic* framework that does not allow for stochastic fluctuations of any kind. In a deterministic model, the dynamics of allele (or genotype) frequencies is governed by the expected values: mutation and recombination rates determine the expected number of mutants or recombinants, and fitness defines the expected number of surviving offspring individuals. In reality, however, the number of offspring of a given individual (and the number of mutants and recombinants) follows a distribution. Altogether, there are three possible reasons why an individual may have many or few offspring:

- *Good or bad genes*: the heritable genotype determines the distribution for the number of surviving offspring. Fitness, in particular, is the expected value of this distribution and determines the allele frequency change due to natural selection.
- *Good or bad environment*: the offspring distribution and the fitness value may also depend on non-heritable ecological factors, such as temperature or humidity. These factors can be included into a deterministic model with space- or time-dependent fitness values. They can also be stochastic, but typically affect all individuals of the population.
- *Good or bad luck*: the actual number of offspring, given the distribution, will depend on random factors that are not controlled by either the genes nor the external environment: chance events that typically affect single individuals. This gives rise to a stochastic component in the change of allele frequencies: *random genetic drift*.

We are interested in the evolutionary change in the number of individuals that belong to a certain class, given the genotypes and environmental parameters. Because of the law of large numbers, genetic drift can be ignored if and only if the number of individuals in each class tends to infinity (or if the variance of the offspring distribution is zero). Note, however, that genetic drift may be relevant even in infinite populations if the number of individuals in a focal allelic class is small.

2.1 The Wright-Fisher model

The Wright-Fisher model (named after Sewall Wright and Ronald A. Fisher) is the standard population genetic model for genetic drift. We will introduce the model for a single locus in a haploid population of constant size N. Further assumptions are no mutation and no selection (neutral evolution) and discrete generations. The life cycle is as follows:

1. Each individual in the parent generation produces an equal and very large number of gametes (or seeds). In the limit of seed number $\to \infty$, we obtain a so-called *infinite gamete pool*.
2. We sample N individuals from this gamete pool to form the offspring generation.

Sewall Wright, 1889–1988, was an American geneticist. 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. His papers on inbreeding, mating systems, and genetic drift make him a principal founder of theoretical population genetics, along with R.A. Fisher and JBS Haldane. Wright's most eminent contribution to population genetics is his concept of *genetic drift* and his development of mathematical theory combining drift with the other evolutionary forces. He was also the inventor/discoverer of key concepts like the *fitness landscape* and the *inbreeding coefficient* and originated a theory to guide the use of inbreeding and crossbreeding in the improvement of livestock (adapted from Encyclopedia Britannica and Wikipedia).

Obviously, this just corresponds to *multinomial sampling with replacement* directly from the parent generation according to the rule:

- Each individual from the offspring generation picks a parent at random from the previous generation and inherits the genotype of the parent.

Remarks

- Mathematically, the probability for $k_1, \ldots, k_N$ offspring for individual number $1, \ldots, N$ in the parent generation is given by the multinomial distribution with

$$\Pr\Big[k_1, \ldots, k_N \big| \textstyle\sum_i k_i = N\Big] = \frac{N!}{\prod_i k_i! N^N}. \tag{2.1}$$

- The number of offspring of a given parent individual is binomially distributed with parameters $n = N$ (number of trials) and $p = 1/N$ (success probability):

$$\Pr\big[k_1\big] = \binom{N}{k_1} \left(\frac{1}{N}\right)^{k_1} \left(1 - \frac{1}{N}\right)^{N-k_1}.$$

- Under the assumption of *random mating*, a diploid population of size N can be described by the haploid model with size $2N$, if we follow the lines of descent of all gene copies separately. Technically, we need to allow for selfing with probability $1/N$.

- The Wright-Fisher model can easily be extended to non-constant population size $N = N(t)$, simply by taking smaller or larger samples to generate the offspring generation.

- Inclusion of mutation, selection, and migration (population structure) is straightforward, as shown in later sections.

2.2 Consequences of genetic drift

Genetic drift is the process of random changes in allele frequencies in populations. We will study its effects using the Wright-Fisher model. To this end, consider a single locus with two neutral alleles a and A in a diploid population of size N. We thus have a haploid population size (= number of gene copies) of $2N$. We denote the number of A alleles in the population at generation t as n_t and its frequency as $p_t = n_t/2N$. The transition probability from state n_t to state $n_{t+1} \in \{0, 1, \ldots, 2N\}$ is given by

$$\Pr[n_{t+1}|n_t] = \binom{2N}{n_{t+1}} \cdot \left(\frac{n_t}{2N}\right)^{n_{t+1}} \cdot \left(1 - \frac{n_t}{2N}\right)^{2N-n_{t+1}} . \tag{2.2}$$

Some elementary properties of this process are:

1. For the expected number of A alleles, we have $\mathrm{E}[n_{t+1}|n_t] = 2N \cdot \frac{n_t}{2N} = n_t$, and thus

   $$\mathrm{E}[p_{t+1}] = \mathrm{E}[p_t] .$$

   The expected allele frequency is constant. We can also express this in terms of the expected change in allele frequencies as $\mathrm{E}[\delta p_t] = \mathrm{E}[p_{t+1} - p_t] = 0$.

2. For the variance among replicate offspring populations from a founder population with frequency $p_t = n_t/2N$ of the A allele, we obtain: $\mathrm{Var}[n_{t+1}|n_t] = 2Np_t(1 - p_t)$ and thus

   $$V := \mathrm{Var}[p_{t+1}|p_t] = \frac{p_t(1 - p_t)}{2N} .$$

   The variance is largest for $p_t = 1/2$. In terms of allele frequency changes, we also have $\mathrm{Var}[\delta p_t] = \mathrm{Var}[p_{t+1} - p_t] = \mathrm{Var}[p_{t+1}|p_t] = V$.

3. There are two absorbing states of the process: Fixation of the A allele at $p = 1$ and loss of the allele at $p = 0$. We can determine the fixation probability p_{fix} at $p = 1$ as follows. Assume that we start in state $p_0 = i/2N$. Since any process will eventually be absorbed in either $p = 0$ or $p = 1$, we have

   $$\lim_{t\to\infty} \mathrm{E}[p_t] = p_0 = p_{\mathrm{fix}} \cdot 1 + (1 - p_{\mathrm{fix}}) \cdot 0 \quad \Rightarrow \quad p_{\mathrm{fix}} = p_0 .$$

   In particular, the fixation probability of a single new mutation in a population is $p_{\mathrm{fix}} = 1/2N$.

Random genetic drift has consequences for the variance of allele frequencies among and within populations. For the variance among colonies that derive from the same ancestral founder population with allele frequency p_0, we have $V = p_0(1 - p_0)/2N$ after a single generation. After a long time, we get

$$V_\infty = \lim_{t\to\infty} \left(\mathrm{E}[(p_t)^2] - \left(\mathrm{E}[p_t]\right)^2\right) = p_0 - p_0^2 = p_0(1 - p_0) .$$

The variance among populations thus increases to a finite limit. To measure variance within a population, we define the homozygosity F_t and the heterozygosity H_t as follows

$$F_t = p_t^2 + (1-p_t)^2 \quad ; \quad H_t = 1 - F_t = 2p_t(1-p_t)\,.$$

The homozygosity (heterozygosity) is the probability that two randomly drawn individuals carry the same (a different) allelic state, where the same individual may be drawn twice (i.e. with replacement). We obtain the single-step iteration

$$F_t = \frac{1}{2N} + \Big(1 - \frac{1}{2N}\Big) F_{t-1}$$

Indeed, if we take two random alleles (with replacement) from the population in generation t, the probability that we have picked the same allele twice is $1/2N$. If this is not the case, we choose parents for both alleles in the previous generation $t-1$. By definition, the probability that these parents carry the same state is F_{t-1}. From this we get for the heterozygosity

$$H_t = \Big(1 - \frac{1}{2N}\Big) H_{t-1} = \Big(1 - \frac{1}{2N}\Big)^t H_0 \approx H_0 \exp[-t/2N]\,.$$

We see that drift reduces variability within a population and $H_t \to 0$ as $t \to \infty$. The characteristic time for approaching a monomorphic state is given by the (haploid) population size. We can derive the half-life for H_t as follows

$$\frac{H_t}{H_0} \approx \exp[-t_{1/2}/2N] := \frac{1}{2} \quad \Rightarrow \quad t_{1/2} = 2N \log[2] \approx 1.39N.$$

The half-life scales with the population size. Note that heterozygosity and homozygosity (as defined here) should not be confused with the frequency of heterozygotes and homozygotes in a population. Both quantities only coincide under the assumption of random mating. For this reason, some authors (e.g. Charlesworth and Charlesworth 2010) prefer the term *genetic diversity* for H_t.

2.3 Neutral theory

In a pure drift model, genetic variation within a population can only be eliminated, but never created. To obtain even the most basic model for evolution, we need to include mutation as the ultimate source for new variation. These two evolutionary forces, mutation and drift, are the only ingredients of the so-called *neutral theory*, developed by Motoo Kimura in the 50s and 60s. Kimura famously pointed out that models without selection already explain much of the observed patterns of polymorphism within species and divergence between species. Importantly, Kimura did not claim that selection is not important for evolution. It is obvious that purifying selection is responsible for the maintenance of functional important parts of the genome (e.g. in coding regions). However, Kimura claimed that most differences that we see within and among populations are not

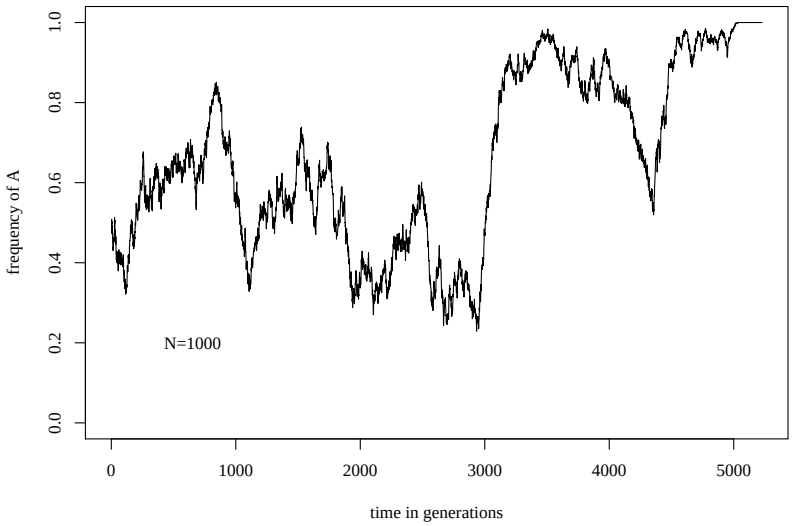

Figure 2.1: Frequency curve of one allele in a Wright-Fisher Model. Population size is $2N = 2000$ and time is given in generations. The initial frequency is 0.5.

influenced by selection. Today, selection is thought to play an important role also for these questions. However, 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`.)

Mutation schemes

There are three widely used schemes to introduce (point) mutations to a model of molecular evolution:

1. With a finite number of alleles, we can define transition probabilities from any allelic state to any other state. For example, there may be k different alleles A_i, $i = 1, \ldots, k$ at a single locus and a mutation probability from A_i to A_j given by μ_{ij}. Mutation according to this scheme is most easily included into the Wright-Fisher model as an additional step on the level of the infinite gamete pool, just like in the deterministic model (1.13). We then obtain the frequencies in the next generations by multinomial sampling with the allele frequencies after mutation.

2. If we take a whole gene as our locus, we get a very large number of possible alleles if we distinguish different amino acid sequences. In particular, back mutation to an ancestral allelic state becomes very unlikely. In this case, it makes sense to assume an effectively infinite number of alleles in an evolutionary model,

   $$A_1 \to A_2 \to A_3 \to \ldots$$

   Usually, a uniform mutation rate u from one allelic state to the next is assumed in the *infinite alleles model*.

3. In the infinite alleles model, we assume that the latest mutation erases all the memory of the previous state. Only the latest state is visible. However, for a stretch of DNA, point mutation rates at a single site (or nucleotide position) are very small. We can thus assume that subsequent point mutations will always happen at different sites and remain visible. This leads to the so-called *infinite sites model* for mutation that is widely applied in molecular evolution. In particular, under the assumptions of the infinite sites model (no "double hits"), we can count the number of mutations that have occurred in a sequenced region – given that we have information about the ancestral sequence.

Predictions from neutral theory

We can easily derive several elementary consequences of neutral theory, given one of the mutation schemes above.

- Under the infinite sites model, new mutations enter a population at a constant rate $2Nu$, where u is the mutation rate per generation and per individual for the locus (stretch of DNA sequence) under consideration. Since any new mutation has a fixation probability of $1/(2N)$, we obtain a neutral substitution rate of

  $$k = 2Nu \cdot \frac{1}{2N} = u\,.$$

Importantly, the rate of neutral evolution is independent of the population size and also holds if $N = N(t)$ changes across generations. As long as the mutation rate u can be assumed to be constant, neutral substitutions occur constant in time. They define a so-called *molecular clock*, which can be used for dating of phylogenetic events.

- For the evolution of the homozygosity F_t or heterozygosity H_t under mutation and drift, we obtain for the infinite alleles model or the infinite sites model

$$F_t = 1 - H_t = (1-u)^2\left(1 - \left(1 - \frac{1}{2N}\right)H_{t-1}\right).$$

In the long term, the population will approach a state where both forces, mutation and drift balance. We thus reach an equilibrium, $H_t = H_{t-1} = H$, with

$$H = \frac{1-(1-u)^2}{1-(1-u)^2(1-1/2N)} = \frac{\Theta(1-u/2)}{\Theta(1-u/2)+(1-u)^2} \approx \frac{\Theta}{\Theta+1}$$

where $\Theta = 4Nu$ is the population mutation parameter. For the special case of the expected *nucleotide diversity*, denoted as $\mathrm{E}[\pi]$, where the focus is on a single nucleotide site, we usually have $\Theta \ll 1$. We can then further approximate

$$\mathrm{E}[\pi] = H_{\text{nucleotide}} \approx \Theta.$$

3 The coalescent

Until now, in our outline 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 current generation (t). This is the classical approach in population genetics and 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 observable data.
- The coalescent leads to very efficient simulation procedures.
- Most importantly, the coalescent allows for an intuitive understanding of patterns in DNA polymorphism data and of how these patterns result from evolutionary processes.

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

The coalescent process describes the genealogy 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

$$p_{c,1} = \Pr[\text{common parent one generation ago}] = \frac{1}{2N}. \tag{3.1}$$

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

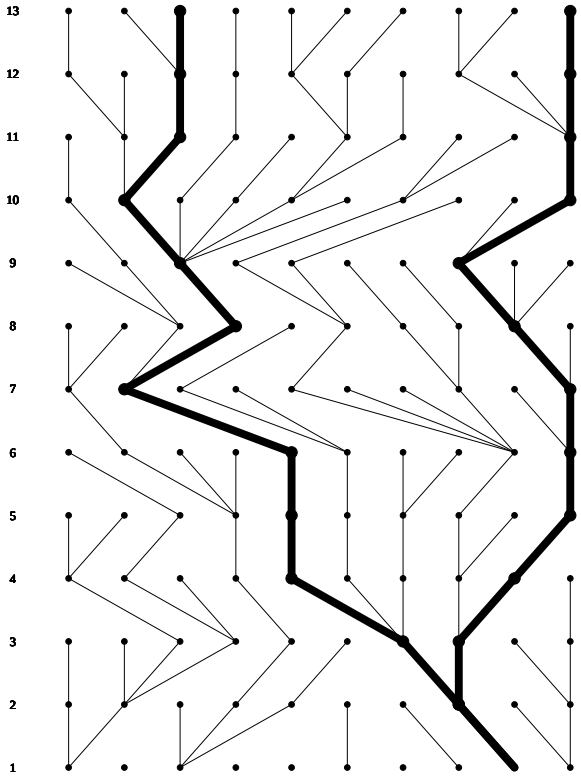

Figure 3.1: The coalescent of two lines in the Wright-Fisher Model

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

$$p_{c,t} = \Pr\left[\begin{array}{l}\text{two lineages coalesce}\\ t \text{ generations ago}\end{array}\right] = \frac{1}{2N}\Big(1-\frac{1}{2N}\Big)^{t-1}.$$

Mathematically, we can describe the *coalescence time* as a random variable that is geometrically distributed with success probability $\frac{1}{2N}$. Figure 3.1 shows an example for the common ancestry like it can be generated by a simulation animator, such as the Wright-Fisher animator on www.coalescent.dk. 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 further back in time 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).

The coalescence perspective is not restricted to a sample of size two but can be applied to any number of individuals. For a sample of size n from the Wright-Fisher model of size

$2N$, the probability of coalescence in a single generation is

$$p_{c,1}^{(n)} = 1 - \left(1 - \frac{1}{2N}\right) \cdot \left(1 - \frac{2}{2N}\right) \cdots \left(1 - \frac{n-1}{2N}\right) = 1 - \prod_{i=1}^{n-1} \left(1 - \frac{i}{2N}\right)$$
$$= \frac{1}{2N} \sum_{i=1}^{n-1} i + \mathcal{O}\left[\left(\frac{n}{N}\right)^2\right] = \frac{1}{2N}\binom{n}{2} + \mathcal{O}\left[\left(\frac{n}{N}\right)^2\right]. \tag{3.2}$$

We can interprete this result as follows. In a sample of size n, there are $\binom{n}{2}$ possible coalescence events between pairs of individuals. If we assume that $n \ll N$, multiple coalescence events in a single generation can be ignored and the leading order term in $p_{c,1}^n$ just accounts for the probability of a single pairwise coalescence event in the sample in the previous generation. Multiple coalescence events and coalescence events of more than two lineages simultaneously (so-called "multiple mergers") only contribute to the error term $\sim \mathcal{O}[N^-2]$, which can be ignored for small samples in a large population. In this approximation, the coalescence probability after t generation in a sample of size n becomes

$$p_{c,t}^{(n)} \approx \frac{1}{2N}\binom{n}{2} \cdot \left(1 - \frac{1}{2N}\binom{n}{2}\right)^{t-1}. \tag{3.3}$$

We can then construct the genealogical history of the sample in a two-step procedure:

1. First, fix the topology of the coalescent tree. I.e., decide (at random), which pairs of genealogical lineages 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.

3.1 Coalescence times

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:

- If X is geometrically distributed with small success probability p and t is large then

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

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

Let t_n be the time until the first coalescence occurs in a smaple of size n. This time is geometrically distributed according to

$$\Pr[t_n > t] = \left[1 - \frac{\binom{n}{2}}{2N}\right]^t = \left[1 - \frac{n(n-1)}{4N}\right]^t. \tag{3.4}$$

The mean waiting time until the first coalescence event is $\mathrm{E}[t_n] = 4N/n(n-1)$ and thus proportional to the population size. It is standard to integrate this dependence into a "coalescent time scale"

$$\tau := \frac{t}{2N}.$$

We can then take the limit $N \to \infty$ to obtain a stochastic process with a continuous time parameter τ. Coalescence times $T_n := t_n/2N$ in this limiting process are distributed like

$$\Pr[T_n > \tau] = \lim_{N\to\infty}\left[1 - \frac{\binom{n}{2}}{2N}\right]^{2N\tau} = \exp\left[-\tau\binom{n}{2}\right]. \tag{3.5}$$

In a sample of size n, the time to the first coalescence is thus exponentially distributed with parameter $\lambda = n(n-1)/2$. The fact that in the coalescent the times are exponentially distributed enables us to derive several important quantities.

- The time to the MRCA,

  $$T_{\mathrm{MRCA}}(n) = \sum_{k=2}^{n} T_k,$$

  is the sum of $n-1$ mutually independent exponentially distributed random variables. Its expectation derives to

  $$\mathrm{E}[T_{\mathrm{MRCA}}(n)] = \sum_{k=2}^{n}\mathrm{E}[T_k] = \sum_{k=2}^{n}\frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{n}\Big). \tag{3.6}$$

  We have $\mathrm{E}[T_{\mathrm{MRCA}}(n)] \to 2$ for large sample sizes $n \to \infty$. Note that $\mathrm{E}[T_{\mathrm{MRCA}}(2)] = 1$, so that in expectation more than half of the total time to the MRCA is needed for the last two ancestral lines to coalesce.

- For the total tree length,

  $$L(n) = \sum_{k=2}^{n} kT_k,$$

  we obtain the expected value

  $$\mathrm{E}[L(n)] = \sum_{k=2}^{n} k\,\mathrm{E}[T_k] = 2\sum_{k=2}^{n}\frac{1}{k-1} = 2\sum_{k=1}^{n-1}\frac{1}{k}. \tag{3.7}$$

  Increasing the sample size will mostly add short twigs to a coalescent tree. As a consequence, also the total branch length

  $$\mathrm{E}[L(n)] \approx 2(\log(n-1) + \gamma) \quad ; \quad \gamma = 0.577216\ldots.$$

  increases only very slowly with the sample size (γ is the Euler constant).

3.2 Polymorphism patterns

In order to generate DNA diversity patterns using the coalescent, we need to add mutations to the process. This can be done according to any of the mutation schemes introduced in section (2.3). Most frequently used is the infinite sites model, which we will discuss in the following.

The key insight for the description of neutral DNA diversity using the coalescent is that neutral mutations do not interfere with the genealogy: *state* (the genotype) and *descent* (the genealogical relationships) are decoupled for neutral evolution. This is easy to see from the time-forward dynamics, since parents carrying different variants of a neutral allele are still equivalent concerning the distribution of their offspring in all future generations. If we want to create a random neutral polymorphism pattern using the coalescent process, we can therefore pick a genealogy first (as described in the previous section) and decide on the state later on. This is done by so-called *mutation dropping*, where mutations are added to all branches of the tree.

For the infinite sites mutation scheme, each mutation hits a new site (and thus leads to a new allele) and all mutations on a genealogy remain visible. If a mutation occurs on a branch of size i in the genealogy of n individuals, it will give rise to a polymorphism with frequency i/n of the derived allele. This means: the mutant allele is seen in i out of n sequences in the sample. Note that we do not need to know the precise time for the origin of the mutations in the genealogy, all that is needed is the total number of mutations that fall on each branch. On genealogical time scales (as opposed to phylogenetic time scales), we can usually assume that the mutation rate u (per haploid individual and generation) is constant.

For a branch of length l, we therefore directly get the number of neutral mutations on this branch by drawing from a Poisson distributed with parameter $2Nlu$. The factor $2N$ accounts for the fact that branch length l is measured on the coalescent time scale (in units of $1/2N$). In particular, the total number of mutations in an entire coalescent tree of length L is Poisson distributed with parameter $2NLu$. Let S be the number of segregating (polymorphic) sites in a sample. Since each polymorphic site corresponds to exactly one mutation on the tree under the infinite sites model, we have

$$\Pr[S=k] = \int_0^\infty \Pr[S=k|\ell] \cdot f_{L(n)}(\ell) d\ell = \int_0^\infty e^{-2N\ell u} \frac{(2N\ell u)^k}{k!} \cdot f_{L(n)}(\ell) d\ell \,.$$

For the expectation that means

$$\begin{aligned} \mathrm{E}[S] &= \sum_{k=0}^{\infty} k \Pr[S=k] = \int_0^\infty \frac{\ell\theta}{2} e^{-\ell\theta/2} \Big(\sum_{k=1}^{\infty} \frac{(\ell\theta/2)^{k-1}}{(k-1)!} \Big) \cdot f_{L(n)}(\ell) d\ell \\ &= \frac{\theta}{2} \int_0^\infty \ell \; f_{L(n)}(\ell) d\ell = \frac{\theta}{2} \mathrm{E}[L(n)] = \theta \sum_{i=1}^{n-1} \frac{1}{i} = a_n \theta \end{aligned} \tag{3.8}$$

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}\,, \tag{3.9}$$

and where

$$\theta = 4Nu$$

is the standard population mutation parameter. Note that the distribution of S does not depend on the coalescent topologies, but only on the distribution of the coalescence times.

The site frequency spectrum

The total number S of polymorphic sites is the simplest so-called *summary statistic* of polymorphism data. There are many more. In particular, we can ask for the number S_i of mutations that are observed in i out of n sequences in the sample. The expected value $\mathrm{E}[S_i]$ can be derived (in a lengthy calculation), with the pleasingly simple result

$$\mathrm{E}[S_i] = \frac{\theta}{i}. \tag{3.10}$$

Together, these numbers define the (expected) *site frequency spectrum* of sample taken from a standard neutral population.

- The frequencies of the expected normalized site frequency spectrum are $p_i = 1/(a_n i)$. They are independent of θ. The characteristic $(1/i)$-shape is a prime indicator of "neutrality".
- We can easily obtain an empirical site frequency spectrum from any polymorphism data. This empirical spectrum can then be compared to the spectrum predicted under neutrality. Note that we need data from many independent (unlinked) loci to observe the *expected* spectrum. For any single locus, the spectrum can differ considerably, because we only have a single coalescent history.
- To determine the size of a given polymorphism in the sample, we need to know the ancestral state at the locus. In practice, this is inferred from a so-called outgroup (usually a single consensus sequence from a closely related sister species). If the ancestral state cannot be determined, we can work with the so-called *folded site frequency spectrum*, with mutation classes $\tilde{S}_i = S_i + S_{n-i}$ for $i < n/2$ and $\tilde{S}_i = S_i$ for $i = n/2$.

3.3 Coalescent and statistics

Coalescent trees show the genealogical relationships between two or more sequences that are drawn from a population. This should not be confounded with a phylogenetic tree that shows the relation of two or more species. Indeed, both "trees" have entirely different roles for the theory of evolution. In phylogenetics, one is usually interested in the one "true

tree" and the parameters of this tree (such as split times) are estimated from data. In contrast, there is no single "true tree" for a set of individuals from a population. Indeed, the genealogy will usually be different for different loci. 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. But even for a single locus, we are usually not able to reconstruct a single "true coalescence tree" and this is not the goal in coalescent studies. Instead, coalescent histories are used as a statistical tool for inferences about an underlying model.

The general idea is as follow. We define an evolutionary model that depends on a number of biological parameters (such as mutation rates, population sizes, selection coefficients). Under this model, we obtain a distribution of coalescent histories and (consequently) a distribution of polymorphism patterns that is predicted under this model. We can then compare measured data with the predicted distribution to make statistical inferences. Usually, there is a twofold goal:

1. to reject (or not) the underlying model. This is true, in particular, for the neutral model as the standard null model of population genetics.

2. to estimate model parameters. Note that the parameters of the coalescent tree (coalescent times, topology) are generally not model parameters. They are "integrated out" in the statistical treatment.

In some easy cases (notably the neutral model), key aspects of the distribution of polymorphism patterns can be obtained analytically using coalescent theory. In many other cases, this is no longer possible. However, even in these cases, the coalescent offers a highly efficient simulation framework that is routinely used in statistical simulation packages.

Estimators for the mutation parameter θ

All population genetic models, whether forward or backward in time, depend on a set of biological parameters that must be estimated from data. In the standard neutral model, there are two such parameters: the mutation rate u and the population size N. However, since both parameters only occur in the combination $\theta = 4Nu$, the population mutation parameter is effectively the only parameter of the model. From our derivation of the expected site frequency spectrum, we easily obtain several estimators for θ. In principle, we can use the total number of mutations of any size class to define an unbiased estimator $\hat{\theta}_i$,

$$\mathrm{E}[S_i] = \frac{\theta}{i} \quad \longrightarrow \quad \hat{\theta}_i := i \cdot S_i \,. \tag{3.11}$$

In practice, widely used estimators are linear combinations across mutations of different size classes. They can be distinguished according to the relative weight that is put on a certain class. The most important ones are the following:

1. *Watterson's estimator*,

$$\hat{\theta}_{\mathrm{W}} := \frac{S}{a_n} = \frac{1}{a_n}\sum_{i=1}^{n-1} S_i = \frac{1}{a_n}\sum_{1\le i\le n/2} \tilde{S}_i\,, \tag{3.12}$$

uses the total number of segregating sites and puts an equal weight on each mutation class. The last equation expresses $\hat{\theta}_{\mathrm{W}}$ in terms of frequencies of the folded spectrum. Remember that the distribution of S – and thus of $\hat{\theta}_{\mathrm{W}}$ – is independent of the coalescent topologies, but only depends on the coalescent times.

2. Let π_{ij} be the number of differences among two sequences i and j from our sample. We have $\mathrm{E}[\pi_{ij}] = \mathrm{E}[S(n=2)] = \theta$. If the sample size is just two, this corresponds to Watterson's estimator. In a larger sample, we can still take the pairwise difference as our basis and average over all $n(n-1)/2$ pairs. This leads to the *diversity-based estimator* (sometimes also called *Tajima's estimator*),

$$\hat{\theta}_{\pi} := \frac{2}{n(n-1)}\sum_{i<j} \pi_{ij}\,. \tag{3.13}$$

We can also express $\hat{\theta}_{\pi}$ in terms of the (folded) frequency spectrum as follows,

$$\hat{\theta}_{\pi} = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)S_i = \binom{n}{2}^{-1}\sum_{1\le i\le n/2} i(n-i)\tilde{S}_i\,. \tag{3.14}$$

Whereas Watterson's estimator weights all frequency classes equally, $\hat{\theta}_{\pi}$ puts the highest weight on classes with an intermediate frequency. In contrast to $\hat{\theta}_{\mathrm{W}}$, it also depends on the distribution of tree topologies. The estimator is often also just written as $\hat{\pi}$.

3. *Fay and Wu's estimator*,

$$\hat{\theta}_{\mathrm{H}} := \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i^2 S_i\,, \tag{3.15}$$

puts a hight weight on mutation classes of the unfolded spectrum with a high frequency of the derived allele. In contrast to the other estimators, it is not a summary statistic of the folded spectrum and thus requires knowledge of the ancestral state.

4. Finally, the *singleton estimator* $\hat{\theta}_{\mathrm{s}}$ uses the singletons of the folded spectrum,

$$\hat{\theta}_{\mathrm{s}} := \frac{n-1}{n}\Big(S_1 + S_{n-1}\Big) = \frac{n-1}{n}\tilde{S}_1\,. \tag{3.16}$$

It has all its weight at both ends of the unfolded spectrum.

Test statistics for neutrality tests

Estimators of any model parameter, such as θ, will only produce meaningful results if the assumptions of the underlying model hold. In our case, we have assumed standard neutral evolution. In addition to the absence of selection, this includes the assumptions of a constant population size and no population structure. But how can we know whether these assumptions do hold (at least approximately) for a given data set? This question asks for a test of the model assumptions. As it turns out, the availability of various different estimators of the same quantity θ is helpful for the construction of such a test.

The key idea is to consider the difference among two different estimators, such as $\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}$. Under standard neutrality, this quantity should be close to zero, whereas significant deviations indicate that the model should be rejected. The most widely used test statistic that is constructed in such a way is *Tajima's D*,

$$D_\mathrm{T} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}]}}\,. \tag{3.17}$$

The denominator of D_T is used for normalization and makes the distribution of the statistic (almost) independent of θ and of the sample size. Tajima has shown that D_T is approximately β-distributed. Today, however, the exact distribution under the standard neutral null model is usually obtained (resp. approximated to arbitrary precision) by computer simulations. For a given significance level α, one can then specify the critical upper and lower bounds for D_T, beyond which the null model should be rejected. Test statistics that are constructed in a similar way are *Fu and Li's D*,

$$D_\mathrm{FL} := \frac{\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}}{\sqrt{\mathrm{Var}[\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}]}} \tag{3.18}$$

and *Fay and Wu's H*,

$$H_\mathrm{FW} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}]}}\,. \tag{3.19}$$

To understand, which kind of deviations from the standard neutral model are picked up by the three summary statistics, it is instructive to consider the contribution of the site frequency classes S_i to the numerator of each statistic. For example, D_T will be negative if we have an excess of very low or very high frequency alleles, whereas it will be positive if many sites segregate at intermediate frequencies.

4 Effective population size

In the previous chapter, we have constructed the coalescent for an idealized Wright-Fisher population. Our assumptions have included the following:

1. neutral evolution with identical offspring distribution for all individuals,
2. a constant population size,
3. no population structure: i.e. offspring choose their parents with equal probability for all individuals from the parent generation,
4. offspring choose their parents independently of each other: as a consequence, the distribution of offspring for each parent is binomial (and approximately Poisson),
5. generations are discrete, individuals are haploid, and there are no separate sexes ...

One may wonder whether such a simplified theory can tell us much about nature. In statistical terms: if we construct a null model under a large number of assumptions, rejecting this null model does not provide us with a lot of information. Indeed, any of the assumptions could have been violated – for most biological populations we even know in advance that several assumption do not hold.

Luckily, the situation is not as bleak as it may look and we can often still use the theory that we have developed. As it turns out, many biological factors can be taken care of by an appropriate adjustment of the model parameters. This leads to the concept of the effective population size.

4.1 The concept

The number of individuals in a natural population is referred to as the *census* population size or *per-capita* population size. *Prima facie*, it seems natural to identify the number of individuals (or individual gene copies) in a Wright-Fisher model with the census population size of a natural population. However, as it turns out, this is usually not appropriate. The point of the Wright-Fisher model (and similar models, like the Moran model) is to capture genetic drift. It should therefore be chosen in such a way that the strength of drift in the natural system is equal to the strength of drift in the model. The idea is to choose the size of an ideal Wright-Fisher population in such a way, that this correspondence holds. The size that is needed is called the *effective* population size. The remaining question is which measure for genetic drift we should use. Unfortunately, there is more than one measure, which leads to some ambiguity in the definition of the effective population size. In general, we use the following philosophy:

> Let $\bullet$ be some measurable quantity that relates to the strength of genetic drift in a population. This can be e.g. the neutral allele frequency variance (or standard deviation) between generations 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 $\bullet$. To be specific, we call N_e the $\bullet$-effective population size.

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 coalescent 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 single-generation *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. If we assume that this probability is the same for all pairs of individuals (no population structure) and constant in time (no demographic changes), we can iterate the single-generation step across generations to obtain the coalescence probability after t generations, $p_{c,t} = p_{c,t}(1-p_{c,1})^{(t-1)}$, as a simple function of $p_{c,1}$. 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 coalescent effective population size

$$N_e^{(c)} = \frac{1}{2p_{c,1}}. \tag{4.1}$$

All coalescent times are directly proportional to this size. One also says that $N_e^{(c)}$ 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

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

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

$$N_e^{(v)} = \frac{p(1-p)}{2\mathrm{Var}[p']}. \tag{4.2}$$

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.

Estimating the effective population size

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

$$N_e^{(c)} = \frac{\mathrm{E}[S_2]}{4u}.$$

In a sample of size n, we can estimate the expected number of pairwise differences to be $\widehat{\mathrm{E}}[S_2] = \hat{\theta}_\pi$ (see (3.13)) and obtain the estimator of $N_e^{(c)}$ from polymorphism data as

$$\hat{N}_e^{(c)} = \frac{\hat{\theta}_\pi}{4u}.$$

A similar estimate can be obtained from Watterson's estimator $\hat{\theta}_\mathrm{W}$, see Eq. (3.12). While the assumption of the infinite sites model is often justified (as long as $4N_e^{(c)}\,u_n \ll 1$, with u_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.

4.2 Factors affecting N_e

Let us now discuss the main factors that influence the effective population size. For simplicity, we will focus on $N_e^{(c)}$. 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 the 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 $\sum_i x_i = 2N$. Then the probability that individual i is the parent of two randomly drawn individuals from the offspring generation is $x_i(x_i-1)/(2N(2N-1))$. Thus, the expected probability for identity by descent of two random offspring individuals is

$$p_{c,1} = \mathrm{E}\Big[\sum_{i=1}^{2N} \frac{x_i(x_i-1)}{2N(2N-1)}\Big] = \sum_{i=1}^{2N} \mathrm{E}\Big[\frac{x_i(x_i-1)}{2N(2N-1)}\Big]. \tag{4.3}$$

With $\mathrm{E}[x_i] = 1$ and $\mathrm{E}[x_i^2] - \big(\mathrm{E}[x_i]\big)^2 = \mathrm{E}[x_i^2] - 1 = \sigma^2$ and the definition (4.1) we arrive at

$$N_e^{(c)} = \frac{1}{2p_{c,1}} = \frac{N-1/2}{\sigma^2} \approx \frac{N}{\sigma^2}. \tag{4.4}$$

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 the 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. Even without selection, this is not necessarily the case. An important example is a population with separate 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, in general, N_f be the number of breeding females and N_m the number of breeding males. 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

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

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

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

The coalescent (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 $N_e^{(c)} = N$.

Fluctuating Population Sizes

Consider the evolution of a population with periodically varying size over a period of T_p generations with values N_0 to N_{T_p-1}. We can ask whether we can describe these fluctuations by an averaged effective population size. Equating the probability that two lineages do *not* coalesce during one period in the fluctuation model with a constant-size model, we obtain

$$\Big(1 - \frac{1}{2N_0}\Big) \cdots \Big(1 - \frac{1}{2N_{T_p-1}}\Big) = \Big(1 - \frac{1}{2N_e}\Big)^{T_p} \tag{4.5}$$

We then have

$$\begin{aligned} 1 - \frac{1}{2N_e} &= \Big[\Big(1 - \frac{1}{2N_0}\Big) \cdots \Big(1 - \frac{1}{2N_{T_p-1}}\Big)\Big]^{1/T_p} \approx \Big[\exp\Big(-\frac{1}{2N_0}\Big) \cdots \exp\Big(-\frac{1}{2N_{T_p-1}}\Big)\Big]^{1/T_p} \\ &= \exp\Big(-\tfrac{1}{2T_p}\Big(\frac{1}{N_0} + \ldots + \frac{1}{N_{T_p-1}}\Big)\Big) \approx 1 - \frac{1}{2T_p}\Big(\frac{1}{N_0} + \ldots + \frac{1}{N_{T_p-1}}\Big) \end{aligned}$$

and get a (coalescent) effective population size of

$$N_e^{(c)} = \frac{1}{2}\frac{1}{\bar{p}_{c,1}} \approx \frac{T_p}{\frac{1}{N_0} + \ldots + \frac{1}{N_{T_p-1}}}.$$

- The effective population size of a fluctuating population is thus 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 $N_e^{(c)} = 14$.

- To justify the use of a constant effective size, we need to make sure that coalescence times (and their distributions) in both models are equivalent. In general, this will be the case if the population runs through many population-size cycles during a typical coalescent time. I.e., we should have

  $$T_p \ll \binom{n}{2}^{-1} N_e^{(c)} = \mathrm{E}[T_n] \tag{4.6}$$

  (in per-generation scaling). This is typically fulfilled, for example, for species with several generations per year and seasonal variation in population size.

- Note that strict periodicity in population sizes is not required for the definition of an average effective size. It is generally sufficient that the population sizes experienced during time periods of $E(T_n)$ are representative of a long-term distribution of population sizes.

From the cases studies in this section, we see that most populations are genetically much smaller than they appear from their census size, increasing the effects of drift. In terms of the coalescent, the change to the effective size means that all coalescent trees are rescaled by a factor N_e/N, but their shape and topology remain the same.

4.3 Larger demographic changes and population structure

We have seen in the previous section that short-term fluctuations in the population size can be subsumed in an effective population size N_e that is the harmonic mean of the population sizes over the period of the fluctuation. This holds as long as this period is short relative to the typical coalescence time. This is no longer true, however, if population sizes change over longer time scales or if there is spatial population structure.

Population growth or decline

If a population experiences long-term growth or decline, there is no equilibrium distribution of population sizes across generations at all, invalidating the concept of an "equilibrium effective size". The simplest model is that of an exponentially growing (or declining) population,

$$N(\tau) = N_0 e^{-\lambda\tau}, \tag{4.7}$$

where λ quantifies the speed of growth and we measure time τ in the backward direction. For the coalescent, this means that two individuals at time τ coalesce in a single generation with probability $1/2N(\tau)$. For a growing population, $N(\tau)$ declines as we go back in time and the frequency of pairwise coalescent events increases. Clearly, this cannot be reproduced by any model with a constant (effective) population size. Graphically, coalescent trees are rescaled by a time-dependent instead of a constant factor. A typical coalescent tree in an expanding population has reduced branch lengths near the root of the tree, as shown in Figure 4.1. Although the shape of the coalescent trees is skewed by such a procedure, their topology is not affected. We can now ask for the effect of such a time-dependent rescaling on the expected summary statistics for the polymorphism pattern.

1. For a growing population, the rescaling will reduce "older" branches near the root of the tree by a larger factor than "young" branches at the leaves. As a consequence, most mutations will fall on branches near the leaves, where they affect only a single individual in the sample. We thus obtain an excess of low-frequency polymorphisms in the site-frequency spectrum relative to the standard neutral model. Looking at the test statistics, we see that Tajima's D_{T} will be negative, while Fay and Wu's H_{FW} will be positive.

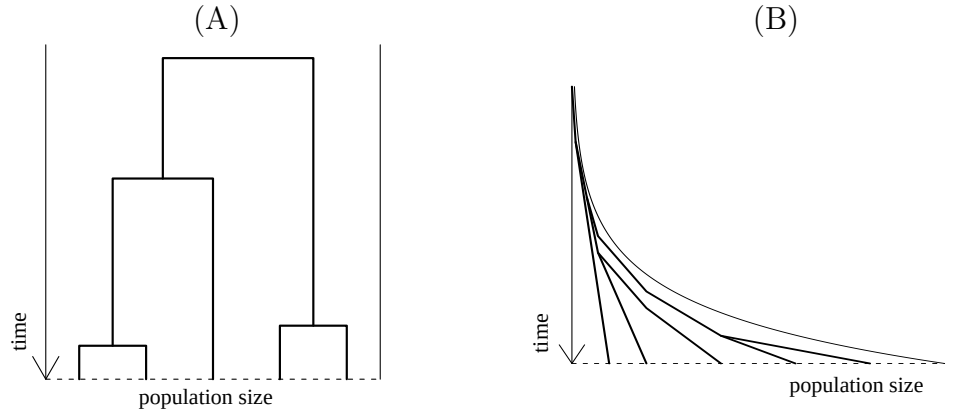

Figure 4.1: (A) The coalescent for a population of constant size and (B) for an expanding population. Real time runs from top to bottom and coalescent time τ from bottom to top.

2. For a shrinking population, the rescaling reduces primarily the branches at the leaves. We typically obtain trees with a deep split, where most of the time during the genealogy is spent for the last two lines to coalesce. Mutations on these branches produce polymorphisms of any size (from 1 to $n-1$) with an equal probability, resulting in a flat site-frequency spectrum. In particular, the number of singletons is reduced relative to the standard neutral model. In contrast to population growth, such a pattern is usually characterized by a positive value of D_{T}.

Population bottlenecks

A complex demographic scenario are so-called bottlenecks, where the population recovers after an intermediate phase with a reduced population size. The consequences of such a demographic history depends on the parameters of the bottleneck in a subtle way. This is seen in two examples in Figure 4.2. On the one hand, a very strong and/or long reduction of the population size can lead to full coalescence of the genealogy with a very high probability during that phase. In this case, the genealogy never "feels" the larger population size further back in time. Consequently, the polymorphism pattern of a very strong bottleneck looks like the one of an expanding population. On the other hand, for a less severe or very short reduction of the population size, two or more lineages will likely survive the bottleneck wthout coalescing. As these lines enter the ancestral phase with large population size, the time to full coalescence at the MRCA can be very long and the polymorphism pattern can even mimic the one of a declining population. For intermediate bottleneck strengths, both of these genealogical scenarios can occur with a high probability. We then get a mix of patterns and a large increase in the *variance* of most summary statistics if we analyze patterns from different loci along a chromosome. This variability of patterns makes it difficult to rule out a bottleneck scenario if we want to infer selection as the cause for a distorted pattern.

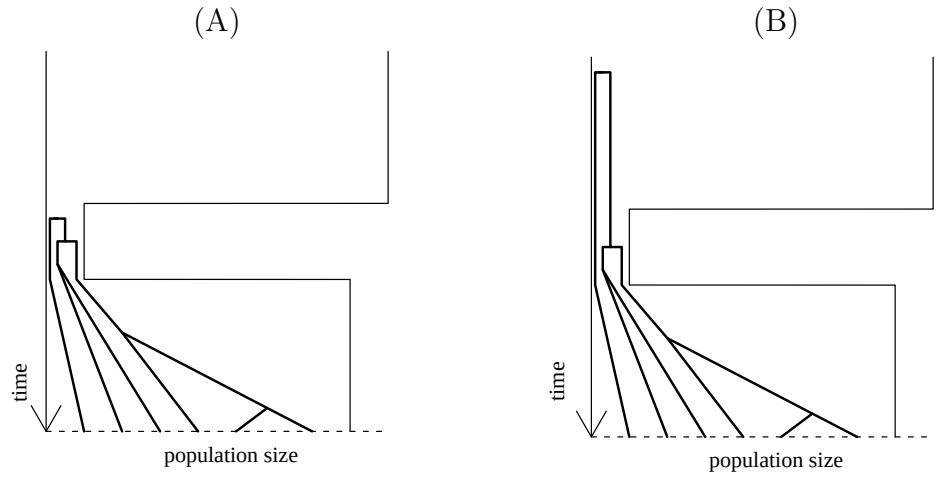

Figure 4.2: Two cases in a bottleneck mode. (A) Only one ancestral line survives the bottleneck. (B) Two or more lines survive which leads to different patterns in observed data.

Population structure

So far, we have assumed that any two individuals from the population can coalesce (have the same ancestor in the previous generation) with the same probability. This is rarely the case in a natural population. Indeed, most populations are structured in the sense that the probabilities for reproduction (forward in time) or coalescence (backward) depend on additional external factors. One of the most important factors to cause population structure is geographic space: individuals that live close to each other are more likely to be closely related than individuals in distant regions. If individuals are no longer "exchangeable" in this sense, this has important consequences on the genealogies.

In a simple scenario, we can think of a population that lives on two islands. The subpopulations on each island are panmictic (i.e., there is no further fine-structure), but individuals on different islands can only trace back to a common ancestor if there is migration in the ancestry of at least one of these individuals. We thus need to consider two types of processes for the construction of genealogies: coalescence and (backward) migration. If N_1 and N_2 is the population size on island 1 and 2, respectively, and n_1 and n_2 the corresponding sample sizes from these islands, these genealogical events occur with the following probabilities

$$p_{c,1}^{(1)} = \binom{n_1}{2}\frac{1}{2N_1} \quad ; \quad p_{c,1}^{(2)} = \binom{n_2}{2}\frac{1}{2N_2}, \tag{4.8}$$

$$p_{m,1}^{1\to 2} = n_1 m_{12} \quad ; \quad p_{m,1}^{21} = n_2 m_{2\to 1}, \tag{4.9}$$

where m_{ij} is the probability that the parent of an individual on island i comes from island j (the so-called backward migration rate). A full mathematical analysis of the resulting genealogies is possible, but is beyond this lecture. Instead, we will focus on two limit cases

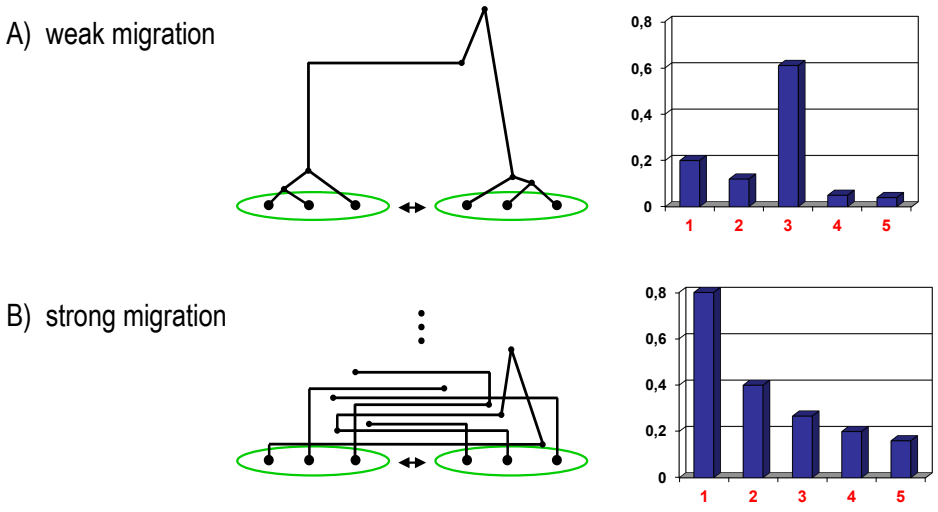

Figure 4.3: Typical coalescent genealogies of a spatially structured population for (A) weak and (B) strong migration and corresponding site-frequency spectra.

of weak or strong migration to understand the qualitative impact of geographic structure on coalescent genealogies.

- If migration is very weak, the subsamples on both islands will most likely coalesce independently before even the first migration event occurs in the genealogy. Only after common ancestors on both islands have been reached, migration will eventually occur on one of the long branches that trace further back in time. As a result, the genealogy of the full sample typically shows a bipartite topology that separates both subsamples. Mutations on the long branches at the root give rise to polymorphisms at intermediate frequencies that show this partition (see Fig. 4.3).

- If migration is very strong, lines of descent typically change back and forth between both island many times before even the first coalescent event occurs. They then reach a so-called migration equilibrium, where each line at a given time is on island 1 or 2 with probability

$$p_1 = \frac{m_{21}}{m_{12}+m_{21}} \quad ; \quad p_1 = \frac{m_{12}}{m_{12}+m_{21}} \, .$$

  Coalescence of any two lineages (independently of the origin of the sampled individual on island 1 or 2) can occur whenever both are on the same island,

$$p_{c,1} = \binom{n_1+n_2}{2}\left(\frac{p_1^2}{2N_1}+\frac{p_2^2}{2N_2}\right).$$

  Since this probability is independent of time, we can use it to define an effective population size

$$N_e = \binom{n_1+n_2}{2}\frac{1}{2p_{c,1}}$$

in the usual way. In the special case that the frequencies in the migration equilibrium are proportional to the population sizes, $p_1 = N_1/(N_1 + N_2)$ etc., the effective size reduces to the census-size, $N_e = N_1 + N_2$. The site-frequency spectrum of a structured population with frequent migration is just the standard neutral one, with the expected total diversity, $\pi = 4N_e u$, adjusted to the (potentially) altered effective size.

- The concept of a structured population can be used in many other contexts. For example, one can define two classes of genes, depending on whether the individual carrier is male or female, or N "islands" with 2 genes each to represent a population of diploids. In both cases, "migration" between these "islands" is strong, and analogous results can be deduced (i.e., we obtain a well-defined coalescent effective size).

We can summarize our findings as follows: In the absence of selection, many biological details concerning the reproductive system and population structure and demography can be fully captured by a single (coalescent-) effective population size N_e. The genealogical structure then conforms to the standard neutral coalescent. In particular, this holds true for all events that are *rapid* on a scale of typical coalescent times, such as rapid fluctuations or strong migration. The two exceptions – other than selection – that lead to altered genealogies are

1. Long-term demographic changes, where population growth leads to star-like genealogies with an expected excess of singletons in the polymorphism pattern, while a shrinking population has a reduced number of singletons. Consequently, Tajima's D is typically negative for population growth and positive for decline. Bottlenecks can give rise to many complicated patterns.

2. Strong population structure (weak migration), which favors coalescent trees with a topology that mirror this structure and typically produce an excess of polymorphism of intermediate frequency (positive Tajima's D).