# Mathematical theory of population genetics

course material, fall 2012

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki
Finland

# Contents

# 1 Elementary population genetics

In this Chapter we give basic concepts of population genetics. This chapter follows largely the book by Bürger (2000).

## 1.1 The Hardy-Weinberg Law

***Two alleles***

Consider a population with discrete and non-overlapping generations. Suppose that individuals mate at random and that either genotype frequencies in both sexes are identical or that every individual has both male and female organs (*monoecious* species, e.g. corn, pine).

Let's denote with $P, 2Q$ and $R$ the relative frequencies (i.e. $P + 2Q + R = 1$) of the three genotypes $A_1A_1, A_1A_2$ and $A_2A_2$. We additionally assume, that $2Q$ is a combined frequency of $A_1A_2$ and $A_2A_1$, and that the population is large so that the relative frequencies of alleles and genotypes can be identified with probability.

Knowing the genotype frequencies in one generation we want to calculate the frequencies of the next generation. This can be done by calculating frequencies of all matings and their offspring. For example, the probability of the mating $A_1A_1 \times A_1A_2$ is $4PQ$, because both genotypes can be either male or female. Table summarizes all the possibilities.

Table 1: mating table

| Mating | Mating prob. | Cond. prob. progeny $(A_1A_1, A_1A_2, A_2A_2)$ |
|---|---|---|
| $A_1A_1 \times A_1A_1$ | $P^2$ | $(1, 0, 0)$ |
| $A_1A_1 \times A_1A_2$ | $4PQ$ | $(\frac{1}{2}, \frac{1}{2}, 0)$ |
| $A_1A_1 \times A_2A_2$ | $2PR$ | $(0, 1, 0)$ |
| $A_1A_2 \times A_1A_2$ | $4Q^2$ | $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ |
| $A_1A_2 \times A_2A_2$ | $4QR$ | $(0, \frac{1}{2}, \frac{1}{2})$ |
| $A_2A_2 \times A_2A_2$ | $R^2$ | $(0, 0, 1)$ |

Therefore, the frequencies of $A_1A_1, A_1A_2$ and $A_2, A_2$ in the next generation are $P' = (P + Q)^2, 2Q' = 2(P + Q)(Q + R)$ and $R' = (Q + R)^2$, respectively (assuming that no evolutionary forces change the frequencies; prime denotes the next generation). Using the derived recurrence equations and the fact that $P + 2Q + R = 1$, we obtain after another generation of random mating

$$P'' = (P' + Q')^2 = (P + Q)^2 = P' \tag{1.1.1}$$
$$Q'' = Q' \tag{1.1.2}$$
$$R'' = R'. \tag{1.1.3}$$

The genotype frequencies are thus maintained in all subsequent generations after one generation of random mating.

Denoting with $p$ and $q = 1 - p$ the frequencies of the alleles $A_1$ and $A_2$, we obtain a relation

$$P' = p^2, \quad 2Q' = 2pq, \quad R' = q^2. \tag{1.1.4}$$

These are the Hardy-Weinberg proportions. The HW law states that after one generation of random mating the genotype frequencies remain the same and can be expressed in terms of the allele frequencies according to (1.1.4). Importantly, as allele frequencies remain constant, this shows that no genetic variability is lost by random mating (**Discuss. Why would genetic variation be lost?**).

***Exercise 2.1*** *Derive the frequency of a genotype $A_iA_j$, where $i, j = 1, 2$, when considering separate sexes (males and females) having different initial frequencies.*

#### *k alleles*

Suppose there are $k$ alleles in the population.

***Exercise 2.2*** *Derive the frequency of a genotype $A_iA_j$, where $i, j = 1 \dots k$.*

## 1.2 Introduction to Quantitative Genetics concepts

To appear soon..

## 1.3 Selection at a single locus

#### *Asexual haploid populations: a demonstration*

By considering a simple genetical underpinning of an individual, this subsection sets up the core idea how selection models are being build. Consider one locus and $k$ alleles $A_1, \dots, A_k$ of an asexually reproducing haploid population (i.e. offspring have the same genotype as the parent, were the genotype of an individual is characterized by the one allele it carries).

Of our focal interest is to study the spread of genes under selection, and consequently, how selection shapes the whole genetic distribution as time goes by. Therefore, we introduce the concept of *fitness*, which is a product of viability (the probability

that an offspring survives to reproductive age; this is the phase of a life-cycle were (ecological) selection acts on the individual), mating success and fertility. In essence, if individuals fitness is greater than one, that is, the individual replaces itself with more than one offspring (individual who carries copies of his genes), then his genes are able to spread to further generations. More precise fitness definitions are given in the section diploid populations.

Let us denote the fitness of individuals carrying allele $A_i$ with $W_i$. The frequency of allele $A_i$ is denoted with $p_i$, such that $\sum_i p_i = 1$. Let $n_i$ denote the number (or more precisely, the density) of individuals of type $A_i$, and let $N = \sum_i n_i$ denote the total population density. Then, the density of $A_i$ in the next generation is $n_i' = W_i n_i$ and the frequency is $p_i' = n_i'/N'$, which implies

$$p_i' = p_i \frac{W_i}{\bar{W}}, \tag{1.3.1}$$

where $\bar{W} = \sum_j W_j p_j$ is the mean fitness of the population. If $p_i(0)$ is the initial frequency of $A_i$, then the solution of (1.3.1) is

$$p_i(t) = \frac{p_i(0) W_i^t}{\sum_j p_j(0) W_j^t}, \qquad i = 1, \ldots, k. \tag{1.3.2}$$

Having now the solution to the dynamical equation (1.3.1), we can investigate the evolutionary outcome (by letting time go to infinity). Suppose, that allele $A_1$ has higher fitness than every other allele (we can for instance relabel the alleles such that $A_1$ is the fittest, i.e. $W_1 > W_j \ \forall j \neq 1$). Then $(W_j/W_1)^t \to 0$ for $j \geq 2$ as $t \to \infty$, and by writing the solution (1.3.2) for $i = 1$ and dividing numerator and denominator with $(W_1)^t$, we have

$$p_1(t) = \frac{p_1(0)(W_1/W_1)^t}{\sum_j p_j(0)(W_j/W_1)^t} = p_1(0) \frac{1}{1 + \sum_{j \geq 2} p_j(0)(W_j/W_1)^t} \tag{1.3.3}$$

which goes to 1 as $t \to \infty$ provided $0 < p_1(0)$. Therefore, the fittest allele $A_1$ will go to fixation and all the other alleles will be lost.

### *Diploid populations*

Let us consider diploid sexually reproducing population with discrete and non-overlapping generations. Assume that population mates at random, and that genotype frequencies are the same in both sexes (considering either monoecious individuals, or dioecious with the same viabilities in both sexes and the same sex ratio in all matings). Consider one (autosomal) locus with $k$ alleles $A_1, \ldots, A_k$, and denote with $P_{ij}$ the ordered frequency of a newborn offspring (zygote) $A_i A_j$. Frequency $P_{ij}$ is called an ordered frequency, when the order of the indices matters (in general, $P_{ij} \neq P_{ji}$). Since genotype frequencies are the same in both sexes, we have that the frequency of heterozygotes $A_i A_j$ is $P_{ij} + P_{ji} = 2P_{ij}$. The frequency of the allele $A_i$ is thus

$$p_i = \sum_j P_{ij}. \tag{1.3.4}$$

Since mating is random, the alleles in the "mating" pool are combined at random to form zygotes and we have that the genotype frequencies are in Hardy-Weinberg proportions (see section 1.1). Denote with $V_{ij}$ the viability of an individual $A_iA_j$. The frequency of type $A_iA_j$ after (ecological) selection (i.e. before mating) is thus

$$P^*_{ij} = \frac{V_{ij}P_{ij}}{\bar{V}} = \frac{V_{ij}p_ip_j}{\bar{V}}, \tag{1.3.5}$$

where

$$\bar{V} = \sum_{i,j} V_{ij}P_{ij} = \sum_{i,j} V_{ij}p_ip_j = \sum_i V_ip_i \tag{1.3.6}$$

is the mean viability and

$$V_i = \sum_j V_{ij}p_j. \tag{1.3.7}$$

The frequency of allele $A_i$ after selection is hence $p^*_i = \sum_j P^*_{ij} = V_ip_i/\bar{V}$. Because of the random mating assumptions (in particular, all the matings are equally likely and result in producing the same number of offspring), the allele frequency in the next generation $p'_i$ among zygotes is also $p^*_i$ (see section 1.1). This results in the *selection equation*

$$p'_i = p_i\frac{V_i}{\bar{V}} \qquad \text{for } i = 1, \ldots, k, \tag{1.3.8}$$

which describes the evolution of allele frequencies at a single autosomal locus in diploid populations when the individuals face selective forces.

Note, that the dynamical equation can be expressed in terms of the viabilities $V_{ij}$. This is because all the matings are assumed equally likely and that if successful they result in the same number of offspring. Indeed, if $V_{ij}$ is multiplied by the same constant (e.g. the product of the mating success and the expected number of offspring) the equation (1.3.8) remains the same. Now, it is often convenient to introduce a scaling that, for example, sets the highest fitness value to 1 (by dividing all the fitnesses by the highest fitness value). Therefore, instead of representing the selection in terms of viabilities or *absolute fitness* (i.e. expected number of progeny of individuals of a given genotype; previously called plainly fitness), we give it in terms of *relative fitness* $W_{ij}$ which is defined as a ratio of the absolute fitness of an individual $A_iA_j$ a3nd the absolute fitness of a reference individual. Therefore, we may write the selection equation as

$$p'_i = p_i\frac{W_i}{\bar{W}} \qquad \text{for } i = 1, \ldots, k, \tag{1.3.9}$$

where $W_i = \sum_j W_{ij} p_j$ is called the *marginal fitness* of the allele $A_i$ (and where $W_{ij}$ is the relative fitness of $A_iA_j$) .

***Exercise 3.1***. The selection equation (1.3.8) can also be derived by explicitly writing out the mating process (using for example a mating table as in the section 1.1) and by assuming equal mating probability and expected number of offspring for all the mating pairs. Mating is assumed to happen after the phase of selection and hence the frequency of $A_iA_j$ during the mating season is given by (1.3.5).

Let us denote with $P^*_{kl}Q_{ij,kl}$ the probability that a female (male) $A_iA_j$ mates with a male (female) of type $A_kA_l$ (for example, $Q_{ij,kl}$ may denote the probability of mating given that individual of type $A_iA_j$ has encountered $A_kA_l$; however, in general, the interpretation of $Q$ can be more complex, and it is only the multiplication $P^*Q$ which gives a probability). Further, let us denote with $F_{ij,kl}$ the expected number of offspring of a couple $A_iA_j$, $A_kA_l$ and denote with $R_{ij,kl\to mn}$ the (Mendelian) probability that parents $A_iA_j$ and $A_kA_l$ produce offspring of type $A_mA_n$. Then, for example, the expected number of offspring produced by a female $A_iA_j$ with a male $A_kA_l$ is simply $P^*_{kl}Q_{ij,kl}F_{ij,kl}$.

Also, note that $Q_{ij,kl} = Q_{ij,lk}$ (doesn't matter from which parent the allele is inherited, applies also to $F_{ij,kl}$), but in general $Q_{ij,kl} \neq Q_{kl,ij}$.

(a) What is the expected number of offspring of type $A_mA_n$ produced by a female $A_iA_j$ with a male $A_kA_l$?

(b) What is the total expected number of offspring produced by a female $A_iA_j$?

(c) What is the total expected number of offspring of type $A_mA_n$ produced by a female $A_iA_j$?

(d) What is the frequency of genotype $A_mA_n$ in the next generation? (Remember to normalize so that frequencies add up to 1 !)

Now, consider two alleles $A_1$ and $A_2$.

(e) Write out the genotype frequencies of the next generation, $P'_{11}, 2P'_{12}, P'_{22}$.

(f) Suppose that all the encounters are equally likely to end up in mating and that all matings result in equal number of offspring, that is, suppose that $Q = Q_{ij,kl}$ and $F = F_{ij,kl}$ for all $i, j, k, l$. Show that the selection equation (1.3.8) is recovered.

***Two alleles***

Let us consider two alleles $A_1$ and $A_2$, with frequencies $p_1 = p, p_2 = 1-p$, respectively. Suppose that the relative fitnesses are

$$W_{11} = 1, \quad W_{12} = 1 - hs \quad \text{and} \quad W_{22} = 1 - s, \tag{1.3.10}$$

where $s$ says how strong is the selective disadvantage of homozygotes $A_2A_2$ compared to the homozygotes $A_1A_1$, and $h$ describes the degree of dominance. To be continued.

***Protected coexistence of two alleles***

In this section we give a sufficient condition for the coexistence of two alleles, and use selection models as an example. Note, however, that the notion of protected coexistence is general and is not restricted two this section.

Allele $A_i$ $(i = 1, 2)$ is said to *protected* if it can increase in frequency when rare. Let's formalize this concept. In the Appendix, we give conditions under which the stability of an equilibrium in a discrete-time dynamical system can be determined by finding the dominant eigenvalue. Then, if the dominant eigenvalue $\lambda_D$ of the Jacobian of the system evaluated at the equilibrium satisfy

$$|\lambda_D| > 1 \tag{1.3.11}$$

the equilibrium is unstable and if

$$|\lambda_D| < 1 \tag{1.3.12}$$

the equilibrium is asymptotically stable. Applying this, we say that allele $A_i$ is protected if $|\lambda_D| > 1$ evaluated at $p_i = 0$, where $i = 1, 2$.

As the selection equation (1.3.9) with two alleles $A_1$ and $A_2$ can be reduced to a one-dimensional system we have that allele $A_1$ is protected when

$$\left[\frac{\partial p'}{\partial p}\right]_{p=0} = \left[\frac{W_1}{\bar{W}}\right]_{p=0} + 0 = \frac{W_{12}}{W_{22}} > 1, \tag{1.3.13}$$

where $p = p_1$. Similarly, allele $A_2$ is protected if $\frac{W_{12}}{W_{11}} > 1$. Now, if both alleles can increase in frequency when rare (i.e. they can not get extinct unless some exterior factors not described in the system cause the extinction), then the alleles must coexist in the interior of the population state space $(0 < p < 1)$. If this is the case, alleles are said to be in *protected coexistence*.

***Exercise 4.2***. Suppose that the viability selection is frequency-dependent, $V_{ij} = V_{ij}(p)$, $i, j = 1, 2$. What is the condition for protected coexistence of $A_1$ and $A_2$?

## 1.4 Migration and selection

This section is based on Chapter 6.2. in Nagylaki (1992) as edited in Bürger (2010).

### *The migration-selection model*

We assume that population is subdivided into $M$ demes (i.e. isolated subpopulations). Within each deme selection acts through differential viabilities. After selection adults migrate, and after migration random mating occurs within each deme.

We consider a single locus with $k$ alleles $A_1, \dots, A_k$. Throughout, we use letters $i, j$ to denote alleles, and greek letters $\alpha, \beta$ to denote demes. We write $G = \{1, \dots, M\}$ for the set of all demes. We denote with $p_{i,\alpha}$ the frequency of allele $A_i$ in deme $\alpha$, so that we have

$$\sum_i p_{i,\alpha} = 1 \tag{1.4.1}$$

for every $\alpha \in G$. As each deme might experience different environmental conditions and hence selection may vary among demes, the viability $W_{ij,\alpha}$ of an $A_iA_j$ individual in deme $\alpha$ may depend on $\alpha$. The (marginal) viability of allele $A_i$ in deme $\alpha$ and the mean fitness of $\alpha$ are

$$W_{i,\alpha} = \sum_j W_{ij,\alpha} p_{j,\alpha} \text{ and } \bar{W}_\alpha = \sum_{i,j} W_{ij,\alpha} p_{i,\alpha} p_{j,\alpha}, \tag{1.4.2}$$

respectively.

Let us now describe migration. Let $\tilde{m}_{\alpha\beta}$ denote the probability that an individual in deme $\alpha$ migrates to deme $\beta$, and let $m_{\alpha\beta}$ denote the probability that an individual in deme $\alpha$ immigrated from deme $\beta$. The $M \times M$ matrices

$$\tilde{\Gamma} = (\tilde{m}_{\alpha\beta}) \text{ and } \Gamma = (m_{\alpha\beta}) \tag{1.4.3}$$

are called the *forward* and *backward* matrices, respectively. Both matrices are *stochastic*, i.e. they are non-negative and satisfy

$$\sum_\beta \tilde{m}_{\alpha\beta} = 1 \text{ and } \sum_\beta m_{\alpha\beta} = 1 \tag{1.4.4}$$

for every $\alpha$. Given the backward migration matrix and the fact that random mating does not change the allele frequencies, the allele frequencies in the next generation are

$$p'_{i,\alpha} = \sum_\beta m_{\alpha\beta} p^*_{i,\beta}, \tag{1.4.5a}$$

where

$$p^*_{i,\alpha} = p_{i,\alpha} \frac{W_{i,\alpha}}{\bar{W}_\alpha} \tag{1.4.5b}$$

describes the change due to selection alone. Substituting we obtain the *migration-selection equation*

$$p'_{i,\alpha} = \sum_\beta p_{i,\beta} \frac{W_{i,\beta}}{\bar{W}_\beta} m_{\alpha\beta}. \tag{1.4.6}$$

The recursion (1.4.6) requires that the backward migration rates are known. As it might be difficult to obtain the knowledge of an individuals origin, we derive a relation between backward and forward migration rates. To this aim, we describe the life-cycle explicitly. It starts with zygotes on which selection acts (possibly including population regulation, i.e. the control of population size, for example, due to limited size of habitat). After selection adults migrate and there may population regulation after migration. Finally, there is random mating and reproduction within each deme. The respective proportions of zygotes, pre-migration adults, post-migration adults, and post-regulation adults in deme $\alpha$ are $c_\alpha, c^*_\alpha, c^0_\alpha$ and $c'_\alpha$ (i.e. $c$'s give the fraction of individuals in each deme at different stages of the life-cycle, so we have that at each point of the cycle the fractions sum up to 1 e.g. $\sum_\alpha c_\alpha = 1$).

Because no individuals are lost during migration, the following must hold:

$$c^0_\beta = \sum_\alpha c^*_\alpha \tilde{m}_{\alpha\beta} \tag{1.4.7a}$$

$$c^*_\alpha = \sum_\beta c^0_\beta m_{\beta\alpha}. \tag{1.4.7b}$$

The joint probability that an individual is in deme $\alpha$ and migrates to deme $\beta$ can be expressed in terms of the forward and backward migration rates as

$$c^*_\alpha \tilde{m}_{\alpha\beta} = c^0_\beta m_{\beta\alpha}. \tag{1.4.8}$$

Substituting (1.4.7a) into (1.4.8), we obtain the desired relation between the forward and backward migration rates

$$m_{\beta\alpha} = \frac{c^*_\alpha \tilde{m}_{\alpha\beta}}{\sum_\gamma c^*_\gamma \tilde{m}_{\gamma\beta}} \tag{1.4.9}$$

Therefore, if $\tilde{\Gamma}$ is given, an *Ansatz* for the vector $c^* = (c^*_1, \ldots, c^*_M)^T$ in terms of $c = (c_1, \ldots, c_M)^T$ is needed to compute $\Gamma$ (as well as a hyphotesis for the variation, if any, of $c$).

Two frequently used assumptions are the following.

1) *Soft selection.* The fraction of adults in every deme is fixed, i.e.,

$$c^*_\alpha = c_\alpha \quad \text{for every} \quad \alpha \in G. \tag{1.4.10}$$

This may be a good approximation if the population is regulated within each deme, e.g. because individuals compete for resources locally.

2) *Hard selection.* Following Dempster (1955), the fraction of adults will be proportional to mean fitness in the deme if the total population size is regulated. This is defined by

$$c_\alpha^* = c_\alpha \frac{\bar{W}_\alpha}{\bar{W}}, \tag{1.4.11}$$

where

$$\bar{W} = \sum_\alpha c_\alpha \bar{W}_\alpha \tag{1.4.12}$$

is the mean fitness of the total population.

These two assumptions are at the extremes of a broad spectrum of possibilities. Soft selection will apply to plants; for animals many schemes are possible.

A migration pattern that does not change deme proportions ($c_\alpha^0 = c_\alpha^*$) is called *conservative.*

## 1.5 Mutation and selection

This section is based on Chapters I.6 and III in Bürger (2010).

Natural selection and mutation are two central factors guiding biological evolution: mutation generates the genetics variability upon which selection can act. The relation between these two processes is the topic of this section.

### *Pure mutation model*

Firstly, we assume all mutations to be neutral, i.e. all genotypes have the same fitness. Let us consider $k$ alleles $A_1, \ldots, A_k$ and label their frequencies with $p_1, \ldots, p_k$.

For $i \neq j$ we denote the probability that an $A_i$ has an $A_j$ as an offspring by the mutation rate $\mu_{ij}$. We shall use the convention $\mu_{ii} = 0$. Then the probability that $A_i$ allele does not mutate is $1 - \sum_j \mu_{ij}$, and $\mu_{ji}$ gives the probability that $A_j$ gives rise to a mutant $A_i$. Therefore, the frequency of $A_i$ in the next generation is

$$p_i' = p_i(1 - \sum_j \mu_{ij}) + \sum_j p_j \mu_{ji}. \tag{1.5.1}$$

Equation (1.5.1) is called the *pure mutation equation.*

**Result** (Bürger 2010) *Equation (1.5.1) has a unique equilibrium if all mutation rates are positive, and the convergence to this equilibrium occurs at a geometric rate.*

Let's demonstrate this result for the case of two alleles.
*Example:* Consider two alleles $A_1$ and $A_2$ and denote the frequency of $A_1$ with $p$ and the mutation rates with $\mu_{12} = \mu$ and $\mu_{21} = \nu$. The recursion (1.5.1) reduces to

$$p' = p(1 - \mu) + (1 - p)\nu = p(1 - \mu - \nu) + \nu. \tag{1.5.2}$$

The equilibrium ($p' = p$) is then

$$\hat{p} = \frac{\nu}{\mu + \nu}, \tag{1.5.3}$$

and it exists (i.e is biologically meaningful) when $\mu, \nu > 0$. Recursion (1.5.2) can be solved (a nice little exercise), and using (1.5.3) its solution can be expressed as

$$p(t) - \hat{p} = (p_0 - \hat{p})(1 - \mu - \nu)^t, \tag{1.5.4}$$

where $p_0 = p(0)$ is the initial frequency of $A_1$. This shows, that for any $p_0$ the solution converges to the equilibrium at a geometric rate, but it's slow, because $\mu+\nu$ is typically very small.

### Mutation-selection equation for haploid populations

In this section we consider a mutation model for haploid populations when selection is taken into account.

The frequency of $A_i$ after selection is $p_i^* = p_i \frac{V_i}{\bar{V}}$, where $\bar{V} = \sum_j V_j p_j$. As multiplying $V_i$ with a constant doesn't change the frequencies, we will write $p_i^* = p_i \frac{W_i}{\overline{W}}$, $\overline{W} = \sum_j W_j p_j$, where $W_i$ denotes the relative fitness. After selection reproduction and mutation occurs and by substituting $p^*$ to (1.5.1) we obtain the *mutation-selection equation*

$$p_i' = p_i \frac{W_i}{\overline{W}} + \frac{1}{\overline{W}} \sum_j (p_j W_j \mu_{ji} - p_i W_i \mu_{ij}). \tag{1.5.5}$$

Note that this equation applies also for diploid populations when $W_i$ denotes the marginal fitness of allele $A_i$ (instead of fitness of genotype $A_i$), and when $\overline{W}$ defines the mean fitness for diploid populations.

It is convenient to transform (1.5.5) into a matrix form. Let us define the $k \times k$ mutation matrix $\tilde{\mathcal{U}} = (\tilde{u}_{ij})$ by

$$\tilde{u}_{ij} = \begin{cases} 1 - \sum_l \mu_{il}, & i = j, \\ \mu_{ji}, & i \neq j, \end{cases} \tag{1.5.6}$$

and the mutation-selection matrix $\mathcal{C} = (c_{ij})$ by

$$c_{ij} = \tilde{u}_{ij} W_j. \tag{1.5.7}$$

Let $\mathbf{p} = (p_1, \ldots, p_k)^T$, then $(\mathcal{C}\mathbf{p})_i = \sum_j c_{ij} p_j$ and

$$\bar{c} = \sum_i (\mathcal{C}\mathbf{p})_i = \sum_i (\sum_j \tilde{u}_{ij} W_j p_j) \tag{1.5.8}$$

$$= \sum_i (\tilde{u}_{i1} W_1 p_1 + \tilde{u}_{i2} W_2 p_2 + \cdots + \tilde{u}_{ik} W_k p_k) \tag{1.5.9}$$

$$= W_1 p_1 \sum_i \tilde{u}_{i1} + W_2 p_2 \sum_i \tilde{u}_{i2} + \cdots + W_k p_k \sum_i \tilde{u}_{ik} = \overline{W}. \tag{1.5.10}$$

Equation (1.5.5) can be rewritten as

$$\mathbf{p}' = \frac{1}{\bar{c}} \mathcal{C}\mathbf{p}. \tag{1.5.11}$$

The state space (i.e. space of allele frequencies) is the simplex

$$S_k = \{\mathbf{p} = (p_1, \ldots, p_k) \in \mathbb{R}^k : \sum_i p_i = 1, p_i \leq 0, i = 1, \ldots, k\}. \tag{1.5.12}$$

Note that $S_k$ is a $(k-1)$ dimensional convex subset of $\mathbb{R}^k$.

Observing that $\mathbf{n}(t) = \mathcal{C}^t \mathbf{n}(0)$ is the solution of $\mathbf{n}' = \mathcal{C}\mathbf{n}$, and $\mathbf{p}(t) = \mathbf{n}(t)/\sum_i n_i(t)$, it follows immediately that (1.5.5) has the explicit solution

$$\mathbf{p}(t) = \frac{\mathcal{C}^t \mathbf{p}_0}{\sum_i (\mathcal{C}^t \mathbf{p}_0)_i}, \tag{1.5.13}$$

where $\mathbf{p}_0 = \mathbf{p}(0) \in S_k$ is the initial frequency distribution.

**Result** (Moran 1976) *If the matrix $\mathcal{C}$ defined in (1.5.7) is primitive, then the mutation-selection dynamics (1.5.5) admits unique equilibrium, $\hat{\mathbf{p}}$, that satisfies $\hat{p}_i > 0$ for every $i$. This equilibrium is the unique solution of*

$$\hat{\overline{W}} \hat{\mathbf{p}} = \mathcal{C}\hat{\mathbf{p}}, \tag{1.5.14}$$

*where $\hat{\overline{W}} = \sum_i W_i \hat{p}_i$ is the equilibrium mean fitness, and it is globally asymptotically stable.*

### Mutation-selection equation for diploid populations

Assume random mating. As mentioned in the previous section, the mutation-selection equation for diploid populations has a similar form as for haploid populations, i.e.

$$p_i' = p_i \frac{W_i}{\overline{W}} + \frac{1}{\overline{W}} \sum_j (p_j W_j \mu_{ji} - p_i W_i \mu_{ij}), \tag{1.5.15}$$

where $W_i = \sum_j W_{ij} p_j$ is the marginal fitness of the allele $A_i$ and $\overline{W} = \sum_{ij} W_{ij} p_i p_j$ is the mean fitness.

***The case of two alleles***. Suppose that the relative fitness values of the genotypes $A_1A_1, A_1A_2, A_2A_2$ are $W_{11} = 1, W_{12} = 1-hs, W_{22} = 1-s$. We denote the frequency of $A_1$ with $p$. For the mutation rates we write $\mu = \mu_{12}$ and $\nu = \mu_{21}$, and we assume that $\mu + \nu < 1$ (this is biologically reasonable, as the mutation rates are usually $\ll 1$). The equilibria can be calculated from (1.5.15) by setting $p = p'$, and solving the resulting polynomial of the third order (therefore there might be up to three solutions).

The equilibria solutions have simple expressions only in special cases. We restrict our attention to the case with no back mutations from the deleterious allele $A_2$ to $A_1$, that is, we set $\nu = 0$. It is convenient to give the precise formulas in terms of $q = 1-p$. Because $A_1$ can't arise by mutation ($\nu = 0$), then if the population consists only of $A_1$ it will always remain so, and hence $\hat{q}^{(0)} = 1$ is always an equilibrium. Since $\nu = 0$ the above mentioned polynomial reduces to a second order polynomial which, if $4\mu/s \leq 1$, has the following solutions in $[0, 1]$ (Bürger 1983):

$$\hat{q}^{(1)} = \frac{h(1+\mu)}{2(2h-1)}\left[1 - \sqrt{1 - \frac{4\mu(2h-1)}{(1+\mu)^2 h^2 s}}\right] \quad \text{if } h \neq \frac{1}{2} \tag{1.5.16a}$$

$$\hat{q}^{(1)} = \frac{2\mu}{s(1+\mu)} \quad \text{if } h = \frac{1}{2} \tag{1.5.16b}$$

and

$$\hat{q}^{(2)} = \frac{h(1+\mu)}{2(2h-1)}\left[1 + \sqrt{1 - \frac{4\mu(2h-1)}{(1+\mu)^2 h^2 s}}\right] \quad \text{if } h_c < h, \tag{1.5.17}$$

where

$$h_c = \frac{1-\mu/s}{1-\mu}. \tag{1.5.18}$$

If $h < h_c$, then $\hat{q}^{(2)} > 1$ and hence biologically not meaningful. In this case ($h < h_c$), the equilibrium $\hat{q}^{(1)}$ is globally asymptotically stable. If $h > h_c$, then three equilibria coexist. They satisfy $0 < \hat{q}^{(1)} < \hat{q}^{(2)} \hat{q}^{(0)} = 1$, and $\hat{q}^{(1)}$ and $\hat{q}^{(0)}$ are asymptotically stable whereas $\hat{q}^{(2)}$ is unstable.

Note that for $h_c \leq h \leq 1$ the pure selection equation model has only one globally stable boundary equilibrium $\hat{q} = 0$, but, the introduction of mutation, however weak, leads to two stable and one unstable equilibria, and in particular, the possible coexistence of $A_1$ and $A_2$. Furthermore, note that the diploid mutation-selection dynamics may be qualitatively different from the haploid dynamics, since in the haploid case, for example, allele $A_2$ is always protected when $\mu > 0, \nu = 0$ (See Exercise 6.3).

Majority of mutations are (slightly) deleterious, and therefore in general, mutations decrease the mean fitness of a population. This decrease is called the *mutation load*.

## 1.6 Two loci

This section is based on Chapters I.6 and II in Bürger (2010).

### 1.6.1 Recombination

Consider a diploid randomly mating population with two loci $\mathcal{A}$ and $\mathcal{B}$, each having an arbitrary number of alleles. Let $p_i$ denote the frequency of $A_i$ in $\mathcal{A}$, and $q_j$ the frequency of $B_j$ in $\mathcal{B}$. The frequency of a gamete $A_iB_j$ we denote with $P_{ij}$, so that we have $p_i = \sum_j P_{ij}$ and $q_j = \sum_i P_{ij}$.

If genes (alleles in two different loci) are associated randomly in a gamete in the sense that

$$P_{ij} = p_i q_j, \quad \forall i, j \tag{1.6.1}$$

is the frequency of $A_iB_j$, then the population is said to be in *gametic equilibrium* (often also called linkage equilibrium). Otherwise, population is in *gametic disequilibrium*. When genes tend to be inherited together due to a close location on a chromosome there is said to be a *linkage*. The gametes individuals produce are said to be of *parental type*, if they contain the same alleles as the gametes the individuals themselves are composed of, and *recombinant types* if the gametes produced are a mix of alleles from both of the parents. For example, if an individual got from his mother $A_1B_1$ and from father $A_2B_2$, then the gametes $A_1B_1$ and $A_2B_2$ produced are called parental types, and $A_2B_1$ and $A_1B_2$ are called recombinant types.

Let us suppose that the proportion of recombinant gametes is $r$, which we call *recombination fraction* (or rate or frequency), and the proportion of parental type is $1-r$. When $r = 0$ loci are completely linked and when $r = \frac{1}{2}$ loci are unlinked and the segregation is independent (typically when loci are on different chromosomes). Thus, in general, recombination fraction satisfies $0 \leq r \leq \frac{1}{2}$ (if $r > \frac{1}{2}$ then there are more recombinant than parental types, which happens only under very special circumstances).

Let us investigate how the gametic frequencies change from generation to generation under random mating and with no evolutionary processes (no mutation, selection etc.). Using the notion of $r$, we have

$$P'_{ij} = (1 - r)P_{ij} + rp_iq_j \tag{1.6.2}$$

for all $i, j$. Thus, gene frequencies are conserved (Exercise), but the gamete frequencies are not (unless $r = 0$ or unless the population is in gametic equilibrium).

Gametic disequilibrium (GD) between alleles $A_i$ and $B_j$ is measured by the parameter

$$D_{ij} = P_{ij} - p_iq_j. \tag{1.6.3}$$

From (1.6.2), (1.6.3) and from the fact that allele frequencies are preserved we get

$$D'_{ij} = P'_{ij} - p'_i q'_j = (1-r)P_{ij} + rp_i q_j - p'_i q'_j \tag{1.6.4}$$
$$= P_{ij} - p'_i q'_j - rD_{ij} \tag{1.6.5}$$
$$= (1-r)D_{ij}, \tag{1.6.6}$$

and hence

$$D_{ij}(t) = (1-r)^t D_{ij}(t). \tag{1.6.7}$$

Therefore, unless $r = 0$, gamete disequilibrium decay at the geometric rate $1 - r$ and gametic equilibrium is approached without oscillation. Note, that even if $r = \frac{1}{2}$ the GE is not reach immediately. Geiringer (1944) showed that the above result holds also for more than two loci.

***Two alleles.*** Let us label the frequencies of $A_1B_1, A_1B_2, A_2B_1, A_2B_2$ with $x_1, x_2, x_2, x_4$, respectively. Let $D$ denote the difference between the frequency of coupling genotypes $A_1B_1/A_2B_2$, and repulsion genotypes $A_1B_2/A_2B_1$, i.e.

$$D = x_1x_4 - x_2x_3. \tag{1.6.8}$$

We have that $D = D_{11} = -D_{12} = -D_{21} = D_{22}$ (Exercise). Thus, the recursion equations for the gamete frequencies (1.6.2) may be rewritten as

$$x'_1 = x_1 - rD \tag{1.6.9}$$
$$x'_2 = x_2 + rD \tag{1.6.10}$$
$$x'_3 = x_3 + rD \tag{1.6.11}$$
$$x'_4 = x_4 - rD \tag{1.6.12}$$

### 1.6.2 Selection model

Consider two alleles at each locus. Let the fitness of zygotes made up of gametes $i$ and $j$ be designated by $W_{ij} (= W_{ji})$, $i, j = 1, 2$. Assume, that there is no position effect, i.e. coupling and repulsion heterozygotes ($A_1B_1/A_2B_2$ and $A_1B_2/A_2B_1$) have the same fitness $W_{14} = W_{23}$. Then the fitnesses can be expressed by the single locus genotypes in the form of a $3 \times 3$ matrix:

$$\begin{array}{c} \\ A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{array} \begin{array}{c} \begin{array}{ccc} B_1B_1 & B_1B_2 & B_2B_2 \end{array} \\ \begin{pmatrix} W_{11} & W_{12} & W_{22} \\ W_{13} & W_{14} & W_{24} \\ W_{33} & W_{34} & W_{44} \end{pmatrix} \end{array} \tag{1.6.13}$$

Each gamete $i$ has a *marginal* fitness $W_i$ defined by averaging over all genotypes containing $i$, i.e.

$$W_i = \sum_{j=1}^{4} W_{ij}x_j, \quad i = 1, \ldots, 4. \tag{1.6.14}$$

The mean fitness of the population is

$$\overline{W} = \sum_{ij} W_{ij} x_i x_j = \sum_i W_i x_i. \tag{1.6.15}$$

Applying (1.6.9) we get the *two allele - two loci selection equation*

$$x_i' = \frac{1}{\overline{W}} \left[ x_i W_i - \eta_i r W_{14} D \right], \quad i = 1, \ldots, 4, \tag{1.6.16}$$

where $\eta_1 = \eta_4 = 1, \eta_2 = \eta_3 = -1$ and $D = x_1 x_4 - x_2 x_3$.

# 2 Introduction to Coalescent Theory

This section is based on Wakeley (2009) and the references within.

Let us first see how coalescent arises in the context of the most commonly applied population genetics models, the Wright-Fisher model and the Moran model. It will be useful to first give forward-time descriptions, and to derive some focal properties.

## 2.1 Random genetic drift in Wright-Fisher and Moran models

In both models, we will make the following assumptions.

- The population size $N$ is constant (when considering haploids this gives $N$ copies of the genome, when diploids then $2N$ copies of the genome)
- All individuals are equally fit
- The population has no geographical or social structure (this implies for example random mating)
- The genes (or sequences) in the population are not recombining

Obviously, above assumption are never met in nature, however, this idealized population is a very useful starting point to build more realistic situations.

### 2.1.1 The Wright-Fisher model

This model was introduced by Fisher (1930) and Wright (1931). In its simplest version, and in addition to the above assumptions, it is assumed that the generations are discrete and non-overlapping, and that the individuals are either haploids or diploids and monoecious (in this case we just replace $N$ with $2N$). For convenience, we will consider haploids.

Finally, and most importantly, WF-model assumes that the genes in generation $t+1$, where $t = 0, 1, \ldots$, are obtained from the parental genes in generation $t$ with replacement (this is a good approximation when individuals are assumed to produce many gametes, so that the proportions of alleles remain constant). After reproduction all the adults die and a new generation begins. To put it simply, consider a box (labeled $t$) with $N$ balls of different colors (each different color represents, say, a different allele). There is another box (a next generation box, labeled $t+1$), which we fill by randomly choosing a ball from the box labeled $t$. After randomly choosing a ball, we make a copy of the ball (represents the production of an offspring) and put one ball to the new box labeled $t+1$ and the other ball we put back to the box $t$. This experiment we repeat $N$ times, so that the next generation box labeled $t+1$ has $N$ balls (population size remains constant).

**Recall some elementary concepts and results from probability theory:**

Consider a discrete random variable (r.v.) $X$ that can assume values $x_1, x_2, \ldots$. The probability function is given by $\Pr\{X = x_i\}$, and by definition

$$\sum_{i=1}^{\infty} \Pr\{X = x_i\} = 1. \tag{2.1.1}$$

The *expected value* of a r.v. $X$ is its expected average over the entire distribution,

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \Pr\{X = x_i\}. \tag{2.1.2}$$

The *variance* of $X$ is

$$\mathrm{Var}[X] = \sum_{i=1}^{\infty} (x_i - \mathrm{E}[X])^2 \Pr\{X = x_i\}. \tag{2.1.3}$$

Variance is often easier to calculate using the formula

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2. \tag{2.1.4}$$

---

Here are some useful rules, where $c$ is a constant:

$$\mathrm{E}[c] = c \tag{2.1.5}$$
$$\mathrm{E}[cX] = c\mathrm{E}[X] \tag{2.1.6}$$
$$\mathrm{E}[X + c] = \mathrm{E}[X] + c \tag{2.1.7}$$
$$\mathrm{Var}[c] = 0 \tag{2.1.8}$$
$$\mathrm{Var}[cX] = c^2\mathrm{Var}[X] \tag{2.1.9}$$
$$\mathrm{Var}[X + c] = \mathrm{Var}[X]. \tag{2.1.10}$$

**Bernoulli trials and binomial distribution:**

*Definition.* A Bernoulli trial is a random experiment in which there are only two possible outcomes - *success* and *failure*.

Examples of a Bernoulli trial are coin tossing, and, a reproduction event where random individual (in a population with two alleles $A_1, A_2$) will pass on one of the two alleles to the next generation.

A Bernoulli r.v. $X$ takes the values 0 and 1 and

$$\Pr\{X = 1\} = p \tag{2.1.11}$$
$$\Pr\{X = 0\} = 1 - p. \tag{2.1.12}$$

(for example, r.v. $X$ is the number of $A_1$ alleles passed on to one offspring by a random individual in a population of two alleles $A_1$ and $A_2$ with frequencies $p$ and $1 - p$, respectively.)

Binomial experiment. Consider a following random experiment:

- The experiment consists of $n$ Bernoulli trials – each trial has two possible outcomes labelled as success and failure
- The trials are independent
- The probability of success in each trial is a constant $p$

*Definition.* The random variable $Y$ that counts the number of successes, $k$, in $n$ trials is said to have a binomial distribution with parameters $n$ and $p$, written Binomial$(n, p)$ (or Binomial$(k; n, p)$= Bin$(n, p)$= B$(n, p)$ ).

The probability mass function of a binomial r.v. $Y$ with parameters $n$ and $p$ is

$$f(k) = \Pr\{Y = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{2.1.13}$$

$\binom{n}{k}$ counts the number of outcomes that include exactly $k$ successes and $n-k$ failures.

Consider two alleles $A_1$ and $A_2$. Let $i$ be the number of copies of allele $A_1$, so that $N - i$ is the number of copies of $A_2$. The frequency of $A_1$ is $p = i/N$ and of $A_2$ is $1 - p$. This gives

$$\pi_{ij} = \binom{N}{j} p^j (1-p)^{N-j}, \quad j = 0, 1, \ldots, N \tag{2.1.14}$$

for the probability that a gene with $i$ copies in the present generation is found in $j$

copies in the next. The $\pi_{ij}$ are called the transition probabilities and the matrix $(\pi_{ij})$ is the *transition matrix* of the associated Markov chain.

Markov chain is a "memoryless" random process: the next step depends only on the current state and not on the sequence of events that preceded it, i.e.

$$\Pr\{X_{i+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i\} = \Pr\{X_{i+1} = x | X_i = x_i\}. \quad (2.1.15)$$

If $K_t$ gives the number of $A_1$ in generation $t$ (say $i$ copies in a population of $N$ genes, as above,) then $K_{t+1}$ is a binomial random variable with parameters $N$ and $p = i/N$ (and we write $K_{t+1} \sim \text{Bin}(N, p)$).

Knowledge of the transition matrix $(\pi_{ij})$ allows one to calculate the probability distribution of $K_t$ for every $t$ if the (in case of not knowing the exact value then the probability distribution of the) initial state $K_0$ is known, because

$$\Pr\{K_{t+1} = j\} = \sum_{i=0}^{N} \Pr\{K_t = i\}\pi_{ij}. \quad (2.1.16)$$

Sums of random variables. If a r.v. $Y$ is a sum of r.v. $X_i$, $i = 1, \ldots, k$, i.e. $Y = X_1 + X_2 + \cdots + X_k = \sum_{i=1}^{k} = X_k$, then

$$\text{E}[Y] = \sum_{i=1}^{k} \text{E}[X_i] \quad (2.1.17)$$

$$\text{Var}[Y] = \sum_{i=1}^{k} \text{Var}[X_i] + \sum_{i=1}^{k} \sum_{j \neq i} \text{Cov}[X_i, X_j]. \quad (2.1.18)$$

If $X_i$ are independent of another, then $\text{Cov}[X_i, X_j] = 0$ and

$$\text{Var}[Y] = \sum_{i=1}^{k} \text{Var}[X_i]. \quad (2.1.19)$$

Let $X_t^m$ be a r.v. of a Bernoulli trial number $m$, $m = 1, \ldots, N$, so that it takes values 0 and 1. $X_t^m = 0$ means that at generation $t$ the $m$th offspring doesn't get $A_1$ from the parent generation $t-1$, and $X_t^m = 1$ that it does. Lets denote with $p$ the frequency of $A_1$ in the parent generation (the probability that $X_t^m = 1$). The mean and variance

of each $X_t^m$ are

$$\mathrm{E}[X_t^m] = \sum_{j=0}^{1} x_j \mathrm{Pr}\{X_t^m = x_j\} = 0 \cdot (1-p) + 1 \cdot p = p \tag{2.1.20}$$

$$\mathrm{Var}[X_t^m] = p(1-p) \tag{2.1.21}$$

We have $K_t = \sum_{m=1}^{N} X_t^m$. Let $K_0 = i$ (the number of $A_1$ at $t = 0$) so that $p_0 = i/N$. Then

$$\mathrm{E}[K_1] = \sum_{m=1}^{N} \mathrm{E}[X_1^m] = \sum_{m=1}^{N} p_0 = Np_0 \tag{2.1.22}$$

$$\mathrm{Var}[K_1] = \sum_{m=1}^{N} \mathrm{Var}[X_1^m] = \sum_{m=1}^{N} p_0(1-p_0) = Np_0(1-p_0). \tag{2.1.23}$$

In fact, if we don't know the initial state $K_0$ with certainty (we only know its distribution), we get $\mathrm{E}[K_1] = \mathrm{E}[K_0]$. It can also be shown that (Exercise)

$$\mathrm{E}[K_{t+1}] = \mathrm{E}[K_t] = \cdots = \mathrm{E}[K_0]. \tag{2.1.24}$$

*Definition.* The *heterozygosity* of a population is defined to be the probability that two randomly sampled gene copies are different.

For a randomly mating diploid, this is the probability that an individual is a heterozygote).

Let $p_0$ be the frequency of $A_1$ at generation $t = 0$ (as above). The heterozygosity at $t = 0$ is

$$H_0 = 2p_0(1-p_0). \tag{2.1.25}$$

Le the r.v. $\mathbb{P}_1$ represent the frequency of $A_1$ at $t = 1$. Then

$$H_1 = 2\mathbb{P}_1(1-\mathbb{P}_1). \tag{2.1.26}$$

Note, that $H_1$ will vary depending on the (random) realization of the process of genetic drift described by (2.1.14). On average

$$\mathrm{E}[H_1] = H_0(1 - \frac{1}{N}) \tag{2.1.27}$$

(Exercise). We get

$$\mathrm{E}[H_t] = H_0(1 - \frac{1}{N})^t. \tag{2.1.28}$$

The random genetic drift hence eliminates all the heterozygosity from the population ($\mathrm{E}[H_t] \to 0$ as $t$ goes to infinity). This implies that one of the alleles becomes fixed (and the other will go extinct). The decrease of heterozygosity is a common measure

of genetic drift, and we say that the drift occurs in the Wright-Fisher model at rate $\frac{1}{N}$ (unit of time is a generation). For large $N$ the heterozygosity decreases exponentially, since

$$\mathrm{E}[H_t] = H_0(1 - \frac{1}{N})^t \approx H_0 e^{-\frac{t}{N}}. \tag{2.1.29}$$

### 2.1.2 The Moran model

This model was introduced by Moran (1958, 1962). It contrasts the model of Wright-Fisher by assuming overlapping generations. Also, we consider only haploid individuals.

In the Moran model, at times $t = 0, 1, 2, \ldots$, a random individual is chosen for reproduction and a random individual is chosen for death. These might be the same individuals or not (in some versions of a Moran model they are not allowed to be the same).

Let the population size be $N$, and let there be $i$ copies of $A_1$ and $N - i$ copies of $A_2$. The number of copies $j$ of $A_1$ in the next generation can assume only three possible values: $i - 1, i, i + 1$. Using $p = i/N$ gives

$$\pi_{ij} = \begin{cases} p(1-p) & \text{if } j = i+1 \\ (1-p)p & \text{if } j = i-1 \\ p^2 + (1-p)^2 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

For example, the probability that $i$ increases is equal to the probability an $A_2$ is chosen to die and $A_1$ to reproduce.

If $K_0 = i$, then with $i = Np_0$ :

$$\mathrm{E}[K_1] = (Np_0 - 1) \cdot p_0(1-p_0) + Np_0 \cdot (p_0^2 + (1-p_0)^2) + \tag{2.1.30}$$

$$+ (Np_0 + 1) \cdot p_0(1-p_0) = Np_0 \tag{2.1.31}$$

$$\mathrm{Var}[K_1] = 2p_0(1-p_0) \tag{2.1.32}$$

(Exercise). Moreover, having $H_0 = 2p_0(1-p_0)$, then

$$\mathrm{E}[H_1] = H_0(1 - \frac{2}{N^2}) \tag{2.1.33}$$

and hence

$$\mathrm{E}[H_t] = H_0(1 - \frac{2}{N^2})^t. \tag{2.1.34}$$

For large $N$ we have

$$\mathrm{E}[H_t] \approx H_0 e^{-\frac{2t}{N^2}}. \tag{2.1.35}$$

The rate of genetic drift is then $2/N^2$. Note, however, that the time units have a different interpretation than in the WF-model. In WF at each $t$ all the adults die, in

Moran only one. To make the time comparable, we define the generation in Moran model as $Nt$ ($N$ death events as in WF), and by scaling time $\tau = t/N$ we obtain

$$\mathrm{E}[H_t] \approx H_0 e^{-\frac{2\tau}{N}}. \tag{2.1.36}$$

The Moran model has twice as high rate of genetic drift as WF (later we see why). However, the decay of heterozygosity is in both models exponential.

## 2.2 The standard coalescent model

In this section we derive the simplest statements of the coalescent model first given in Kingman (1982a, 1982b, 1982c). We begin by deriving the ancestral process of the Wirght-Fisher and the Moran model and then generalize the results.

### 2.2.1 Wright-Fisher model derivation

Kingman proved for the WF-model that the coalescent process (precise definition comes later) describes the ancestral genetic process for a sample of fixed size $n$ in the limit as the population size $N \to \infty$. In particular, he showed that the *coalescent times* $T_i$ are (mutually) independent and exponentially distributed. Coalescent time $T_i$ is a random variable which gives the time during which there are exactly $i$ lineages (ancestral to the sample).

**Discrete-time coalescent: sample of two genes ($n$=2).** What is the distribution of the waiting time until the most recent common ancestor (MRCA) of two genes sampled from $N$ genes?

Note that we will talk about haploid individuals, but similar calculations hold for diploid monoecious populations.

The probability that two genes find a common ancestor in the first generation back in time is

$$\frac{1}{N} \tag{2.2.1}$$

and the probability that two genes have different ancestors is $1 - \frac{1}{N}$. Since sampling in different generations is independent from each other, the probability to have a common ancestor $k$ generations ago is

$$(1 - \frac{1}{N})^{k-1} \frac{1}{N}. \tag{2.2.2}$$

Thus, the coalescent time $T_2$ for two genes to find a MRCA is distributed as

$$\Pr\{T_2 = k\} = \left(1 - \frac{1}{N}\right)^{k-1} \frac{1}{N}, \quad k = 1, 2, \ldots \tag{2.2.3}$$

As $T_2$ is geometrically distributed with parameter $p = \frac{1}{N}$, we get for the expected waiting time (until MRCA)

$$\mathrm{E}[T_2] = \frac{1}{p} = \frac{1}{1/N} = N. \tag{2.2.4}$$

(see also Exercises). The expected time until MRCA is the same as number of genes in the population.

**Discrete-time coalescent: sample of $n$ genes.** Assume for the moment that $N$ is not necessarily large. Our aim is to derive an expression for a (transition) probability that in a single generation $i$ lineages are descended from $j$ ancestors. This can be thought of as tossing $i$ balls randomly with replacement into $N$ boxes. Whenever balls end up in the same box, they share an ancestor ("with replacement" refers to the fact that the same individual can be a parent to multiple offspring). The single generation transition probability is

$$G_{i,j} = \underbrace{\frac{N}{N}\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-j+1}{N}}_{j \text{ different ancestors}} \cdot \underbrace{\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}}_{i-j \text{ common ancestors}} \cdot \underbrace{S_i^{(j)}}_{\text{Stirling number}} = \frac{S_i^{(j)} N_{[j]}}{N^i}, \tag{2.2.5}$$

in which $N_{[j]} = N(N-1)\dots(N-j+1)$ and $S_i^{(j)}$ are Striling numbers of the second kind (# of ways $i$ elements can be partitioned into $j$ subsets). Stirling number can be generated recursively using $S_i^{(1)} = 1$ and

$$S_i^{(j)} = S_{i-1}^{(j-1)} + jS_{i-1}^{(j)}, \quad j = 2, 3, \dots, i-1 \tag{2.2.6}$$

and with $S_i^{(i)} = 1$. Useful special case is

$$S_i^{(i-1)} = \binom{i}{2} = \frac{i(i-1)}{2}. \tag{2.2.7}$$

From (2.2.5) we see that all the transitions have a positive probability (for $j = 1, \dots, i$). However, Kingmans coalescent admits only $j = i$ and $j = i-1$, that is, at most two out of the $i$ lineages share a common ancestor. Next we show that this is the case for WF when $N \to \infty$.

The probability that $i$ lineages have $i$ (distinct) ancestors is

$$G_{i,i} = \frac{N}{N}\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-i+1}{N} \tag{2.2.8}$$

$$= (1-\frac{1}{N})(1-\frac{2}{N})\dots(1-\frac{i-1}{N}) \tag{2.2.9}$$

$$= 1 - \sum_{j=1}^{i-1}\frac{j}{N} + \mathcal{O}(\frac{1}{N^2}) \tag{2.2.10}$$

and the probability that $i$ lineages have $i-1$ ancestors is

$$G_{i,i-1} = \frac{N}{N}\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-i+2}{N}\frac{1}{N}\binom{i}{2} \tag{2.2.11}$$

$$= (1-\frac{1}{N})(1-\frac{2}{N})\ldots(1-\frac{i-2}{N})\frac{1}{N}\binom{i}{2} \tag{2.2.12}$$

$$= \sum_{j=1}^{i-1}\frac{j}{N} + \mathcal{O}(\frac{1}{N^2}) \tag{2.2.13}$$

Note, that all other $G_{i,j}$ with $j < i-1$ are $\mathcal{O}(\frac{1}{N^2})$ (Exercise).

Thus, as $N$ becomes larger and larger, the ancestral process for $i$ lineages becomes like a series of Bernoulli trials with a constant probability $G_{i,i-1} = \frac{i(i-1)}{2N}$ each generation of success. Success in this case means that a single pair of lineages coalesces.

In consequence, the probability that two genes out of $i$ genes find a common ancestor $T_i = k$, $k = 1, 2, \ldots$, generations ago is

$$\Pr\{T_i = k\} = (1-\binom{i}{2}\frac{1}{N})^{k-1}\binom{i}{2}\frac{1}{N} \tag{2.2.14}$$

and $T_i$ has approximately a geometric distribution with parameter $\binom{i}{2}\frac{1}{N}$.

We have that times $T_i$ are independent and geometrically distributed. The goal was, however, to show that they are exponentially distributed as $N \to \infty$. To do this, we need to measure time differently.

Suppose that time $t$ is measured in many steps (for example, in $N$ generations instead of just one generation). Now, consider much smaller time-step $\tau = \delta t$, where $\delta$ is a very small number. For a very small $\delta$, $\tau$ can be seen as a continuous time approximation. From $e^{-x} \approx 1-x$, when $x$ is small, we get that $(1-\lambda\delta t)^{t/\rho t} \to e^{-\lambda\delta t\frac{t}{\rho t}}$ as $\delta t \to 0$, where $\lambda$ is a rate s.t. $(1-\lambda\delta t)$ is the probability that event doesn't occur in a small time unit $\delta t$ and $(1-\lambda\delta t)^{t/\delta t}$ is the probability that event doesn't occur during $t$. In the limit the events are exponentially distributed.

**The continuous-time coalescent.** Suppose that we measure time in $N$ generations. Then, for large $N$ one generation can be seen as a continuous time approximation. Setting $t = \tau N$, we get from $G_{i,i-1}$ that $\binom{i}{2}$ is the rate at which coalescent events occur (where the time unit is $N$ generations). Then,

$$\Pr\{T_i > t\} \to e^{-\binom{i}{2}t}, \quad \text{as } N \to \infty, \tag{2.2.15}$$

is the probability that no coalescent event happens during $t$. We get that the probability density function of $T_i$ is

$$f_{T_i}(t) = \binom{i}{2}e^{-\binom{i}{2}t}. \tag{2.2.16}$$

### 2.2.2 Moran model derivation

In Moran model, only two things can happen

- with probability $1/N$ the same individual is chosen to reproduce and to die. As single offspring replaces its parent, a common ancestor event between two lineages is impossible (both within the sample and the whole population)
- with probability $1 - 1/N$ different individuals reproduce and die, thus parent and offspring coexist in the population in the next time-step. Backwards in time, a common ancestor event occurs in the total (!) population.

Thus, for coalescent event to happen among the $i$ lineages, the $i$ lineages need to contain the parent and the offspring (randomly sampled without replacement). Using simple probabilistic rules, a straightforward calculation shows that

$$G_{i,i-1} = (1 - 1/N)\,\frac{i(i-1)}{N(N-1)} = \binom{i}{2}\frac{2}{N^2} \tag{2.2.17}$$

is the probability that a coalescent event happens among the $i$ lineages in one time-step (exercise).

### 2.2.3 The n-coalescent

In this section we give a formal description of the n-coalescent (Kingmans coalescent) and the convergence theorem for exchangeable-type populations (complete description will appear soon, at this moment we just describe the consequences). It is adopted from Kingman (1982) and Möhle (2000). The convergence theorem is found in Kingman (1982, pg. 101, Theorem 1). The Theorem states that in the limit as population size $N$ goes to infinity, the coalescent times $T_i$ are independent and exponentially distributed as

$$f_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}, \quad i = 2, \ldots, n \tag{2.2.18}$$

when time is measured appropriately ($i$ gives the number of lineages at time $t$). For the Theorem to hold we need the variance $\sigma^2 = \mathrm{Var}(v_1)$ to converge to a non-zero limit as $N \to \infty$, where $v_1$ is the number of offsprings of the 1st individual in the parent generation (note that it doesn't matter which parent we look at, they all have the same variance).

The Theorem also gives the correct time-scale for the coalescent times $T_i$: the time of the model at hand has to be scaled by factor $N_e = \frac{N}{\sigma^2}$ to obtain the time-scale of the ancestral process (see below for an example). $N_e$ is called the *coalescent effective size*.

**$N_e$ in Wright-Fisher and Moran models.** Let us first calculate $\sigma^2$ in the Wright-Fisher model. The joint distribution of the numbers of offspring $V_1, V_2, \ldots, V_N$ each

generation of the $N$ individuals is *multinomial* with parameters $N$ and $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$. We get

$$\mathrm{E}[V_i] = Np_i = N\frac{1}{N} = 1 \tag{2.2.19}$$

$$\mathrm{Var}[V_i] = Np_i(1-p_i) = 1 - \frac{1}{N} \tag{2.2.20}$$

$$\mathrm{Cov}[V_i, V_j] = -Np_ip_j = -\frac{1}{N}. \tag{2.2.21}$$

As $N \to \infty$, $\sigma^2 = \mathrm{Var}[V_i] \to 1$. We thus have $N_e = N$, that is, measuring time in $N$ generations the coalescent times are exponentially distributed with parameter $\binom{i}{2}$.

For the Moran model the joint distribution of $V_1, V_2, \ldots, V_N$ is not a well known distribution. We will say that if an individual is not chosen for reproduction nor death it leaves one offspring (i.e. instead of saying that the individual survives until the next generation we say it is replaced by one offspring).

***Exercise.*** Show that in the Moran model $\mathrm{E}[V_i] = 1$ and $\mathrm{Var}[V_i] = \frac{2}{N}(1 - \frac{1}{N})$.

In the Moran model, as $N \to \infty$, $\sigma^2 \to 0$, and hence Kingman's convergence theorem can't be applied. In the previous section we however showed that Moran model does converge to the coalescent. Möhle (2000) provides a more general Theorem that also covers the Moran model.

### 2.2.4 Some properties of coalescent genealogies

Let us denote with

$$T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i \tag{2.2.22}$$

the time to the most recent common ancestor (MRCA) of the entire sample $n$ and with

$$T_{\mathrm{total}} = \sum_{i=2}^{n} iT_i \tag{2.2.23}$$

the total length of all the branches in the genealogy. As $T_{\mathrm{MRCA}}$ and $T_{\mathrm{total}}$ are independent r.v., we have

$$\mathrm{E}[T_{\mathrm{total}}] = \sum_{i=2}^{n} i\mathrm{E}[T_i] = \sum_{i=2}^{n} i\frac{2}{i(i-1)} = 2\sum_{i=1}^{n-1}\frac{1}{i} \tag{2.2.24}$$

and

$$\mathrm{E}[T_{\mathrm{MRCA}}] = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\sum_{i=2}^{n}(\frac{1}{(i-1)} - \frac{1}{i}) = \tag{2.2.25}$$

$$= 2(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{2} - \cdots + \frac{1}{n-1} - \frac{1}{n}) = \tag{2.2.26}$$

$$= 2(1 - \frac{1}{n}) \tag{2.2.27}$$

Let us now give the full probability distributions of $T_{\mathrm{MRCA}}$ and $T_{\mathrm{total}}$. The distribution of $T_{\mathrm{MRCA}}$ is the sum of $n-1$ independent exponential r.v. $T_i$, with parameters $\lambda_i = \binom{i}{2} = \frac{i(i-1)}{2}$ for $2 \leq i \leq n$. Because the parameters $\lambda_i$ take different values for different $i$, we need to take a series of $n-1$ convolutions of $T_i$, and we obtain

$$f_{\sum_{i=2}^{n} T_i}(t) = \sum_{i=2}^{n} \lambda_i e^{-\lambda_i t} \prod_{j=2, j\neq i}^{n} \frac{\lambda_j}{\lambda_j - \lambda_i}. \tag{2.2.28}$$

An exponential r.v. can be rescaled by any constant factor to yield a new exponential r.v. with an appropriately rescaled parameter. E.g. If we wish to measure time in units that are $C$ times longer than the old units, we perform $s = t/C$, so that $t = Cs$ and $dt = Cds$. Then

$$f_S(s)ds = \lambda e^{-\lambda t}dt = \lambda C e^{-\lambda C s}ds \tag{2.2.29}$$

Defining $T_i^* = iT_i$, then $T^*$ also follows an exponential distribution,

$$f_{T_i^*}(t)dt = \binom{i}{2} e^{-\binom{i}{2}t^*}dt^* = \frac{i-1}{2} e^{-\frac{i-1}{2}t}dt \tag{2.2.30}$$

and

$$f_{T_{\mathrm{total}}}(t)dt = \sum_{i=2}^{n} \frac{i-1}{2} e^{-\frac{i-1}{2}t} \prod_{j=2, j\neq i}^{n} \frac{j-1}{j-i}. \tag{2.2.31}$$

### 2.2.5 Human-Neanderthal couples?

In recent years there has been a considerable debate whether ancient humans and neanderthals have interbred, and consequently whether there is Neanderthal DNA left in humans, after the time of divergence of their ancestry about $500,000$ years ago ($T_{\mathrm{MRCA}}$ of humans and Neanderthals). It was suggested that this mixing of populations would have happened when human subpopulation migrated out of Africa $100,000 - 30,000$ years ago and encountered Neanderthals in Western Asia (Krings et al. 1997). Krings et al. sampled mtDNA from a Neanderthal that lived $30,000 - 100,000$ years ago, compared it to the mtDNA of 986 modern humans (and some chimpanzees), and

concluded that Neanderthal sequence falls outside human mtDNA variation (thus concluding that ancient humans and Neanderthals didn't interbreed in this period). Furthermore, they estimated that the time $T_e$ for the MRCA of the sample of 986 is at least four times shorter that the MRCA event $T_r$ of humans and Neanderthals, i.e. $T_r \geq 4T_e$.

As the obtained genealogical tree seems unlikely under the assumption of humans and Neanderthal interbreeding $30,000 - 100,000$ years ago, Nordborg (1998) set out to calculate the *actual probability* of such a genealogical tree under a such null-model (where humans and Neanderthal *did* interbreed randomly). He was after a $\Pr\{\text{tree AND } T_r \geq 4T_e\}$ under the null-model. The key idea is to calculate the probability by conditioning on the number of human mtDNA lineages that existed at time $t_s$ (the age of the Neanderthal) when the Neanderthal sequence joins the remaining lineages.

Let $A_n(t)$ denote the number of lineages at time $t$ in the past of a present day sample $n$. Note, that given $A_n(t)$ the probability of the tree and the probability that $T_r \geq 4T_e$ are independent of one another. Thus,

$$\Pr\{\text{tree AND } T_r \geq T_e\} = \sum_{k=1}^{n} \Pr\{T_r \geq 4T_e|k\}\Pr\{\text{tree}|k\}\Pr\{A_n(t) = k\}. \quad (2.2.32)$$

To calculate (2.2.32), Norborg assumed that the population size is constant s.t. the coalescent effective size is $N_e = 3400$ (females, since mtDNA is inherited through females only), the generation time is 20 years and that the Neanderthal sequence is $30,000 - 100,000$ years old. We thus have that in the coalescent time-scale, $t_s$ is between $1500/N_e \approx 0.44$ and $5000/N_e \approx 1.47$.

Let us denote with $T_{n,i}$ the time for a sample of $n$ lineages to coalesce to $i$ lineages (this equals to the time of $n - i$ coalescent events).

Let us first calculate $\Pr\{T_r \geq 4T_e|k\}$ (the other probabilities will be updated to the lecture notes later). Note that $T_r - t_s = T_{k+1,1}$, $T_e - t_s = T_{k+1,2}$, $T_{k+1,1} - T_{k+1,2} = T_2$. We have,

$$\begin{aligned}
\Pr\{T_r \geq 4T_e|k\} &= \Pr\{T_r - 4T_e \geq 0|k\} && (2.2.33)\\
&= \Pr\{(T_r - t_s) - 4(T_e - t_s) \geq 3t_s|k\} && (2.2.34)\\
&= \Pr\{(T_{k+1,1} - 4T_{k+1,2} \geq 3t_s\} && (2.2.35)\\
&= \Pr\{(T_2 - 3T_{k+1,2} \geq 3t_s\} && (2.2.36)\\
&= \Pr\{(T_2/3 - T_{k+1,2} \geq t_s\} && (2.2.37)
\end{aligned}$$

Let $\tilde{T} = T_2/3 - T_{k+1,2}$. As $T_2/3$ and $T_{k+1,2}$ are r.v. involved in non-overlapping coalescence time-intervals, they are independent. Taking their convolution we have

$$f_{\tilde{T}}(t) = \int_0^\infty f_{T_{k+1,2}}(x) f_{T_2/3}(y + x)dx \quad (2.2.38)$$

and hence

$$\Pr\{T_r \geq 4T_e|k\} = \Pr\{\tilde{T} \geq t_s\} = \int_{t_s}^{\infty} f_{\tilde{T}}(t)dy \tag{2.2.39}$$

$$= \int_{t_s}^{\infty}\int_{0}^{\infty} f_{T_{k+1,2}}(x) f_{T_2/3}(y+x)dxdy. \tag{2.2.40}$$

***Exercise.*** Show that the distribution of $T_2/3$ is exponential with parameter $\lambda = 3$ (see 2.2.18).

# References


[1] Bürger, R. (1983) *On the evolution of dominance modifiers I. A nonlinear analysis.* J. Theor. Biol. 101: 585–598.

[2] Kingman, J.F.C. (1982) *Exchangeability and the evolution of large populations.* In: Koch, G., Spizzichino, F. (Eds.), Exchangeability in Probability and Statistics. North-Holland Publishing Company, Amsterdam, pp. 97–112.

[3] Krings, M., Stone, A., Schmitz, R. W., Kraintiski.,Stoneking, M., Pääbo, S. (1997) *Neanderthal DNA sequences and the origin of modern humans* Cell 90: 19–30.

[4] Levene, H. (1953) *Genetic equilibrium when more than one ecological niche is available.* Am. Nat. 87: 331–333.

[5] Möhle, M. (2000) *Ancestral processes in population genetics–the coalescent.* J. Theor. Biol. 204, 629–638.

[6] Nordborg, M. (1998)*On the probability of Neanderthal ancestry* Am. J. Hum. Genet. 63: 1237–1240.