Chapter 1

General Introduction

1.1 Standfirst

The complexity of our world is striking, thrilling, even frightening sometimes. Looking at the Cambridge dictionary definition, complex is defined as "having many parts connected in ways that are difficult to understand". This definition is directly coming from the word's etymology: the word complex comes from the Latin root *plectere*: to weave, entwine, and from the prefix *com* "with, together". Hence the complexity of a system, arises from the interactions between its parts, with the inherent computational difficulties coming from the combinatorial nature of interactions. In its broadest sense, we call "complex system" : 1) a set of entities 2) that through their interactions 3) forms a unified whole defined in terms of its boundaries 4) that present emerging properties not present at the level of each component.

The first strategy proposed to study complex systems was the use the reductionist approach (for a review see (Rosenberg, 2007)). Popularized by René Descartes, in *De homine* (1662), the author was one of the first to develop a reductionist approach of life, considering schematically a complex system as the sum of its parts in first approximation. However the reductionist method isn't without flaws as (i) hidden variables could act as confounding factors, (ii) several causal factors do not always act additively on a system, and (iii) it does not account for relationships between the system and its environment. A more systemic approach is therefore necessary. Although it is arguable that science studies complex systems for hundreds of years, formal conceptualization underlying the complex system theory dates from the 1970's. The complex system theory can be seen as an alternative paradigm to reductionism; it attempts to explain systems in terms of their constituent parts and the individual interactions among them.

In that sense, living organisms are highly complex: interactions are ubiquitous within and between the different scales of life (from molecules as DNA, cells, tissues, organisms to ecosystems). Even without considering interactions, the number of species (parts of the biosphere system) is tremendous, with an estimated 5 ± 3 million leaving species (Costello et al., 2013). Since Darwin published in 1859 *On the Origin of Species*, the evolutionary biology community tries to understand the origin of this immense diversity through the study of the mechanisms that govern the changes in the heritable characteristics of biological populations over successive generations, *i.e.* the adaptive process. These mechanisms are referred to as *evolutionary forces* and impact the allele frequencies of a population, defined here in its largest sense as a collection of individuals. The evolutionary forces structuring these complex leaving systems are:

- Selection: the differential survival and reproduction of individuals due to differences in phenotypes,
- Drift: the change in the frequency of an existing allele in a population due to random sampling of organisms,
- Mutation: the alteration of the nucleotidic sequence of the genome of an organism during reproduction,
- Migration: the gene flow between individuals from different populations,
- Recombination: Recombination doesn't properly speaking change the allelic frequencies, but their

covariance, or in other words the haplotypes of an organism, thanks to crossover between homologous chromosomes happening during sexual reproduction.

- Mating system can be seen as an additional evolutionary forces: in sexual eukaryotes, the mating system is related to sexual behaviour of individuals, the way gametes are formed and individuals encounter upon fecundation when it exists.

Several fields of investigation (namely Mendelian genetics, population genetics, quantitative genetics and the later evolutionary developmental biology, and genomics) have been integrated in the "Modern Synthesis". However, their theoretical framework rely on different views of the adaptive process. For example, population genetics works bottom-up from the dynamics at single loci, without much focus on the phenotype, whereas quantitative genetics envisions phenotypic adaptation top-down, from the vantage point of the trait (Höllinger et al., 2019). Therefore, they differently explore the parameter spaces of their models. While, we are still far from a holistic and unified view of adaptation, experimental evolution combined with -omic data can help bridging the gap between these two disciplines.

Fundamentals in quantitative and population genetics have been set long before the advent of molecular genetics. Those disciplines can be seen as a field of applied mathematics that describe changes of allelic frequencies and/or phenotypes in populations. In the early 1900's, most of the objects manipulated in population genetics were conceptual entities described by letters, using a mathematical formalism (*e.g.* Box "Some definitions" in 1.2.1.1). I wrote this introduction in an educational perspective in the hope to help other students to navigate among the diversity of models and underlying hypotheses. Personally, I would have loved to be taught these concepts prior to the beginning of this intellectual adventure.

In this introduction we will keep in mind the reading axes: how does evolutionary forces and their interactions govern the adaptive process ? What are the consequences on adaptation of two qualitatively different sources of variation, namely standing genetic variation and *de novo* mutation ?

1.2 Evolutionary forces

1.2.1 Drift

Because populations are of finite size, genetic drift occurs whenever the collection of individuals is randomly sampled, resulting in random change of allele frequencies. Consider a population composed of N diploid individuals randomly mating to produce a new population at the next generation. Some individuals may leave many offspring, while others may leave few because of non genetic factors that can be modelled as a stochastic process. Furthermore, mendelian segregation adds another source of stochasticity. Indeed, an heterozygote Aa can transmit either a a allele or a A allele to its progeny. Altogether, at the next generation, the total number of A alleles will be the result of a binomial sampling $B(2N, p)$, where p is the frequency of the A allele at the parental generation, and $2N$ the number of gametes that form the N diploid individuals. This random change in the number of A in the offspring translates into a change in the probability of drawing the allele A or a at the next generation.

1.2.1.1 Wright-Fisher Model for haploid populations

For the sake of simplicity, let's firstly consider a Wright-Fisher Model, whose assumptions are:

- A constant population size of $N \in \mathbb{N}$ haploid individuals every generation.
- Non-overlapping generations.
- No selection: Generation at time t+1 is formed by randomly choosing N individuals at generation time t, uniformly and with replacement. $t \in \mathbb{N}$
- One locus with two alleles A and a.

Intuitively, each parent gives birth to a large number of offspring but only N individuals randomly survive to form the next generation. Figure 1.1 presents different simulation results for the evolution of the allelic frequency p of the A allele in a Wright-Fisher population under genetic drift. We can extract from this graph some intuitive remarks. An allele submitted to drift is submitted to frequency changes characteristic of a brownian motion. An allele either become fixed or lost. The smaller the population size,

Some definitions

- **Locus** : The position of a sequence of interest on a chromosome, like a gene, an exon, a nucleotide.
- **Gene** : A sequence of nucleotides that encodes the synthesis of an RNA often translated into a protein. A concept in constant redefinition.
- **Allele** : A variant form of a locus/gene.
- **Genotype** : The combination of alleles carried by an individual at a particular locus.
- **Ploidy** : The number of complete sets of chromosomes in a cell.
  Haploid: The number of sets of chromosomes normally found in a gamete.
  Diploid: Diploid cells have two complete sets of chromosomes.
- **Allele frequency** : The ratio of the number of copies of an allele in a population, over the total number of alleles. Consider two alleles A and a at a given locus in a population of size N, with N_A and N_a copies:
  For a haploid population, the frequency of allele A is given by: $f_A = \frac{N_A}{N_A+N_a} = \frac{N_A}{N}$.
  For a diploid population, the frequency of allele A is given by: $f_A = \frac{N_A}{N_A+N_a} = \frac{N_A}{2N}$.
  In all cases allelic frequencies sum up to 1: $f_A + f_a = 1$.
- **Genotype frequency** : The ratio of the number of individuals of one genotype in a population, over the population size. Consider two alleles A and a at a given locus in a population of size N:
  For a haploid population, the genotype frequencies equal the allelic frequency, which greatly simplifies the calculations.
  For a diploid population, three different genotypes are possible for a diallelic locus (AA, Aa,aa), and we have:
  $f_{AA} = \frac{N_{AA}}{N}$; $f_{Aa} = \frac{N_{Aa}}{N}$; $f_{aa} = \frac{N_{aa}}{N}$ with $f_{AA} + f_{Aa} + f_{aa} = 1$. We can compute directly the allelic frequencies from the genotypic frequencies, e.g:

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N} = f_{AA} + \frac{1}{2} f_{Aa}$$

- **Hardy-Weinberg equilibrium** : For a diploid species, the allelic frequencies can be computed from the genotypic frequencies. However we need other hypotheses to estimate the genotypic frequencies from the allelic frequencies. These are:
  - Monoecious diploid population.
  - Infinite population size.
  - Panmixia (i.e. random mating: all individuals are potential partners, gametic equiprobability, Mendelian allelic segregation during meiosis).
  - No selection.
  - No mutation.
  - No migration.
  - Discrete non-overlapping generations.

  Genotypic frequencies can then be computed from allelic frequencies:

Genotype	AA	Aa	aa
Frequency	f_A^2	$2f_A f_a$	f_a^2

  Under these hypotheses, one generation is sufficient to reach the so called Hardy-Weinberg equilibrium, where allele frequencies are constant trough time.

the quicker the loss/fixation of the allele is. The smaller the population size, the higher the variance in frequency change through generations. Let's present a theoretical framework to formalize these intuitions.

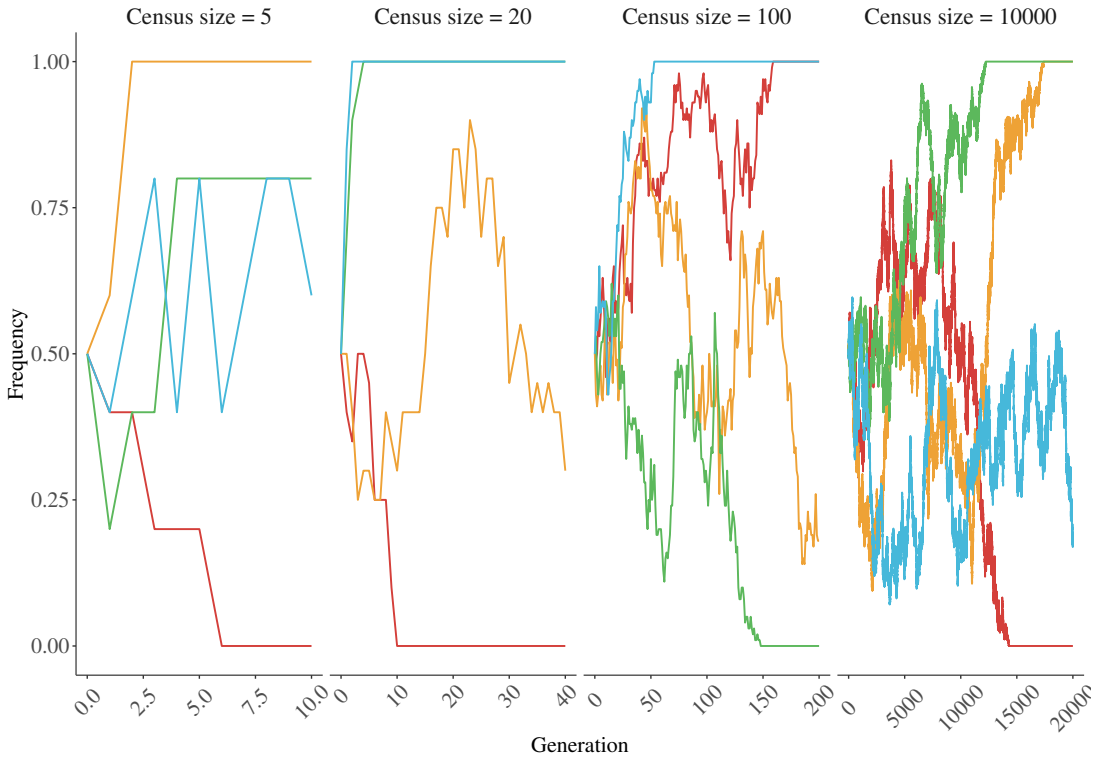

Figure 1.1: **Allele frequency (of allele A) change over generations due to drift in a Wright-Fisher Model.** Each line corresponds to an independent simulation. Note the different x-axis ranges for each different population size.

A Markov process Let X_t denotes the number of individuals of type A in the t'th generation, and $p_t = \frac{X_t}{N}$ be the frequency of allele A. In a Whright-Fisher population, the next generation consists in a random sampling with replacement of N parents. Each parent transmits a copy of its allele to its progeny. Hence, X_{t+1} is a drawing in a Binomial(N,p_t) distribution :

$$\Pr\{X_{t+1} = m | X_t = Np_t\} = \binom{N}{m} p^m (1-p_t)^{N-m}, m \in [\![0, N]\!] \tag{1.1}$$

Note that X_{t+1} does not depend on X_{t-1}, but solely on the value of X_t at the previous generation.

Recall that a Markov chain is a stochastic process noted $(X_t, t \geq 0)$ describing a sequence of possible events in which the probability of each event depends only on the state of the previous event. In other words, a discrete Markov chain is a sequence of random variables $X_1, X_2, X_3, ...$ satisfying the Markov property:

$$\Pr(X_{n+1} | X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_n = x_n) = \Pr(X_{n+1} | X_n = x_n) \tag{1.2}$$

Hence, $(X_t, t \geq 0)$ is a discrete-time Markov chain. As $X_{t+1}|X_t$ follows a binomial distribution, the expected value and the variance of X_{t+1} are:

$$\mathbb{E}(X_{t+1}|X_t) = Np_t = X_t \text{ implying } \mathbb{E}(p_{t+1}|p_t) = p_t \tag{1.3}$$

$$\text{and: } \mathbb{V}(X_{t+1}|X_t) = Np_t(1-p_t) \text{ implying } \mathbb{V}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{N} \tag{1.4}$$

Using the law of total expectation, we obtain:

$$\mathbb{E}(X_{t+1}) = \mathbb{E}(\mathbb{E}(X_{t+1}|X_t)) = \mathbb{E}(N\frac{X_t}{N}) = \mathbb{E}(X_t) \tag{1.5}$$

By induction, we have $\mathbb{E}(X_t) = X_0$, and $\mathbb{E}(p_t) = p_0$.

From these equations, we can see two key properties of genetic drift:

- The expected allele frequency is constant.
- The variance of the allele frequency is inversely proportional to the population size.

Loss of genetic diversity through time Even though allele frequencies do not change on expectation, each particular trajectory is expected to be different, as shown in Figure 1.1. Moreover, as time passes, allele frequencies tend to be closer to fixation or loss, which corresponds to a reduction of the genetic diversity within populations. Genetic diversity can be measured by Nei's index $H_t = 2p_t(1-p_t)$, the probability that two randomly (with replacement) picked individuals contain both alleles in a population a time t.

To see how H_t change over time, we can write

$$\mathbb{E}(H_1) = E(2p_1(1-p_1)) = 2(\mathbb{E}(p_1) - \mathbb{E}(p_1^2)) = 2(\mathbb{E}(p_1) - \mathbb{V}(p_1) - \mathbb{E}(p_1)^2) \tag{1.6}$$

$$= 2(p_0 - \frac{p_0(1-p_0)}{N} - p_0^2) = (1-\frac{1}{N})2(p_0 - p_0^2) \tag{1.7}$$

$$= (1-\frac{1}{N})H_0 \tag{1.8}$$

By induction on t, we have:

$$\mathbb{E}(H_t) = H_0(1-\frac{1}{N})^t \approx H_0\,\mathrm{e}^{-t/N} \textit{ for large } N \tag{1.9}$$

Therefore, the smaller the population, the quicker the loss of genetic diversity will be.

If we set $z_{i,t} = \Pr\{X_t = i\}$, we can write the Markov chain as follows:

$$\overrightarrow{z_{t+1}} = T.\overrightarrow{z_t} \tag{1.10}$$

$$\text{where: } T = \left[\begin{array}{c|cccc} & (0) & (1) & ... & (N) \\ \hline (0) & 1 & 0 & ... & 0 \\ (1) & t_{1,0} & ... & ... & t_{1,N} \\ & ... & ... & ... & ... \\ (N) & 0 & 0 & ... & 1 \end{array}\right] \text{ with: } t_{i,j} = \Pr\{X_{t+1} = j | X_t = i\} \tag{1.11}$$

We can easily see from the transition matrix T that $p = 0$ and $p = 1$ are absorbing states, these states correspond to the fixation of allele A if $p = 1$ and a if $p = 0$.

Therefore, without any other evolutionary forces, whatever the initial conditions (N and p_0), drift leads to the inescapable loss of polymorphism in a population, through the fixation of alleles.

Fixation probabilities Next, we can compute the probability that allele A get fixed. Let $\tau = min\{t : X_t = 0 \text{ or } X_t = N\}$ be the fixation time. The probability of fixation of A writes $\Pr\{X_\tau = N\}$, and the probability of fixation of a writes $\Pr\{X_\tau = 0\}$. As we have seen that fixation always occurs and that the average value of X_t remains constant over time, we have intuitively:
$\mathbb{E}(X_\tau) = 0 \Pr\{X_\tau = 0\} + N \Pr\{X_\tau = N\} = \mathbb{E}(X_1) = X_0$ therefore:

$$\Pr\{X_\tau = N\} = \frac{X_0}{N} = p_0 \tag{1.12}$$

To prove this, we note that since $X_t = X_\tau$ when $t > \tau$, $X_0 = \mathbb{E}(X_t) = \mathbb{E}(X_\tau; \tau \leq t) + \mathbb{E}(X_t; \tau > t)$ where $\mathbb{E}(X; A)$ is short for the expected value of X over the set A.Now, let $t \to \infty$ and use the fact that $|X_t| \leq N$ to conclude that the first term converges to $\mathbb{E}(X_\tau)$ and the second to 0.

An interesting consequence is the rate of fixation of new mutations: a mutation can occur in any of the N individuals with probability μ (the mutation rate). Hence, it goes to fixation with probability $1/N$. So, at each generation, the probability that a mutation arises and ultimately gets fixed is $N \cdot \mu \cdot 1/N = \mu$.

- In simpler terms, the fixation probability of an allele A corresponds to its initial frequency.
- In terms of genealogies, this implies that all alleles in a population descend from one unique allele at generation 0, with a probability p_0 to be of allele A.
- The rate of fixation of a neutral mutation in a population of size N, is the mutation rate μ, and does not depend on N.

Average fixation time Let us now compute the average fixation time $\tau(p_0)$ of an allele that is initially at frequency p_0. If $p_0 = 0$ or $p_0 = 1$, fixation is reached so that $\tau(0) = 0$ and $\tau(1) = 0$. Otherwise ($p_0 \neq \{0, 1\}$), neglecting the probability of fixation within a single generation, the allele frequency at the next generation is p_1, and $\tau(p_0) = 1 + \tau(p_1)$. By conditioning on the value p_1 and using the Markov property, we obtain:

$$\tau(p_0) = 1 + \sum_k \Pr\{p_1 = k/N\}\tau(k/N) \tag{1.13}$$

Therefore $\tau(p)$ is a solution of a certain linear equation. It can be solved when N is small, but become computationally heavy when N becomes large, and an approximation is required. Suppose $\tau(p)$ is continuous, we can expand the function in a Taylor series about p:

$$\tau(p) \approx 1 + \sum_k \Pr\{p_1 = k\}[\tau(p) + \tau'(p)(p_1 - p) + \frac{1}{2}\tau''(p)(p_1 - p)^2] \tag{1.14}$$

$$\tau(p) \approx 1 + \tau(p) + \mathbb{E}(p_1 - p)\tau'(p) + \frac{1}{2}\tau''(p)\,\mathbb{E}((p_1 - p)^2) \tag{1.15}$$

Because $\mathbb{V}(p_{t+1}) = \frac{p_t(1-p_t)}{N}$, and $\mathbb{E}(p_1 - p) = 0$, (1.15) gives,

$$\tau''(p) \approx \frac{-2N}{p(1-p)} \tag{1.16}$$

Solving with the boundary conditions $\tau(0) = \tau(1) = 0$ gives:

$$\tau(p) \approx -2N(p\log(p) + (1-p)\log(1-p)) \tag{1.17}$$

Thus, for the Wright-Fisher model, the expected time to fixation is of order $O(2N)$. But be careful, this formalism called "diffusion equation" is based on a continuous time approximation that is correct if N is large and $p \approx (1-p)$.

Main equation for diploid:

- Heterozygosity loss:

$$\mathbb{E}(H_t) = H_0(1 - \frac{1}{2N})^t \approx H_0\, \mathrm{e}^{-t/2N} \text{ for large } N \tag{1.18}$$

- Time to fixation:

$$\tau(p) \approx -4N(p\log(p) + (1-p)\log(1-p)) \tag{1.19}$$

One can very often extend results for haploid population to diploid one, by considering a population size of 2N. Note that this doesn't hold when interactions among alleles or loci exist.

1.2.1.2 The Coalescent approach in haploid populations

Let us try to have a glimpse of the impact of genetic drift on the genealogies of a population. A famous approach, synthesized under the term "coalescent" or "n-coalescent" was developed by J.F.C. Kingman in his articles: *On the genealogy of large populations* (Kingman, 1982b) and *The coalescent* (Kingman, 1982a). Recall that the ultimate fate of an allele is to be lost or to get fixed because of drift. Suppose that the initial situation is N different alleles. Then, after fixation, all individuals will carry the same allele as one of its ancestors, that has been transmitted from parent to offspring. The coalescent model looks backwards in time and describes the fact that two or more individuals (or alleles) merge/coalesce in the genealogies into a single ancestor through random coalescent events (see figure 1.2).

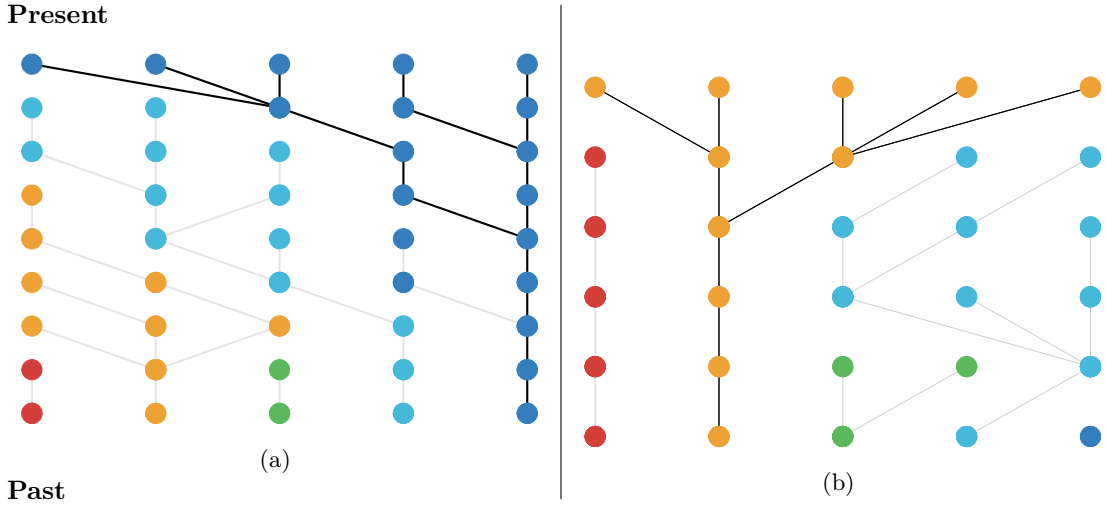

Figure 1.2: **Two simulated coalescent process in a Wright-Fisher population for** $N = 5$. Each dot represents an individual. Each row of dots represents a generation. Present is at the top and past at the bottom. A color represents a genotype. A line represents a kin relationship. The darker line represents the ancestral path of the observed population. In sub-figure (a), the three first dark blue dots share an ancestor at the previous generation. This is called multiple merging. The 2 others dots coalesce 2 generations back in a simple merging process. All five dark blue dots share the same ancestor (or coalesce) 4 generations back. Sub-figure (b) represents a more complicated multiple merging process. The 2 first orange dots coalesce a generation back, in a simple coalescence event, and the three last orange dots also merge a generation back. We have here 2 coalescence events at the same generation, with two and three individuals sharing two different ancestors. The five orange dots coalesce two generations back.

Looking backwards in time, the five individuals of figure 1.2a) coalesce in one generation in three individuals, which coalesce in one generation into two individuals. Hence, all five individuals share

a common ancestor four generations back. Hence a coalescent can be seen as a succession of times $(t_k, t_{k-1}, ..., t_2)$ between coalescence events from the actual population to the common ancestor. In figure 1.2a), we have $(t_5 = 1, t_4 = 0, t_3 = 1, t_2 = 2)$, while in figure 1.2b), we have $(t_5 = 1, t_4 = 0, t_3 = 0, t_2 = 1)$. Note that the individuals that did not leave any progeny at the considered generation are discarded because they don't bring any information.

Two individuals First consider two random individuals from a haploid population of size N. The probability that they share the same parent at time $t-1$ is $\frac{1}{N}$:

Indeed, for the first considered offspring, N parents are possible and N parents are possible for the second, giving N^2 total possibilities but only N possibilities for identical parents: it follows the sought probability: $\frac{N}{N^2} = \frac{1}{N}$.

We can write the probability that two individuals do not share the same parent at generation $t-1$ as: $1 - \frac{1}{N}$. Thus the probability that the common ancestor of these two individuals lived two generations back is $(1 - \frac{1}{N})\frac{1}{N}$. By induction, the probability that these two individuals coalesce n generations back is $(1 - \frac{1}{N})^{n-1}\frac{1}{N}$.

Hence, the time t_2 before the coalesence of two lineages is a random variable that follow a geometric distribution of parameter N^{-1}, with $\mathbb{E}(t_2) = N$ and $\mathbb{V}(t_2) = N(N-1)$.

k individuals Kingman shows that in a sample of k individuals the probability of coalescing at the previous generation is:

$$\frac{k(k-1)}{2}\frac{1}{N} \tag{1.20}$$

Indeed the number of ways to pick two individuals in a sample of k, is: $\binom{k}{2} = \frac{k(k-1)}{2}$ and $\frac{1}{N}$ the probability that they come from the same parent. However, we are neglecting here the fact that more than two individuals can come from the same parent, a process called multiple-merging. For example: in figure 1.2b three individuals come from the same parent a generation back. In figure 1.2a a pair and a triplet of individuals coalesce a generation back. However, as Kingman noted these events occur with probability of order $O(\frac{1}{N^2})$ and can be neglected when N is large.

Therefore, the probability that the k lineages do not coalesce during n generations when N is large is:

$$\approx (1 - \frac{k(k-1)}{2}\frac{1}{N})^n \approx \mathrm{e}^{-\frac{k(k-1)}{2}\frac{n}{N}} \tag{1.21}$$

If we express time in terms of N (the population size) generations, and set $t = \frac{n}{N}$ with n the generation, we can see that, the time t_k during which there are k lineages, converges to an exponential distribution of rate $\frac{k(k-1)}{2}$ and mean $\frac{2}{k(k-1)}$, *i.e.* the expected value of t_k is $\mathbb{E}(t_k) = \frac{2N}{k(k-1)}$, and its variance is $\mathbb{V}(t_k) = \frac{4N^2}{k^2(k-1)^2}$.

Most recent Common Ancestor Looking backward in time, there will be k lineages during t_k, then $k-1$ lineages during t_{k-1}, and so on until the last two lineages coalesce into a particular individual called the Most Recent Common Ancestor (MRCA) of the sample Fig. 1.3.

The time T_{MRCA} to the MRCA can be computed as $T_{MRCA} = \sum_{i=2}^{k} t_i$.

We then have:

$$\mathbb{E}(T_{MRCA}) = \mathbb{E}(\sum_{i=2}^{k} t_i) = \sum_{i=2}^{k} \mathbb{E}(t_i) = \sum_{i=2}^{k} \frac{2N}{i(i-1)} = 2N \sum_{i=2}^{k} \frac{1}{i} - \frac{1}{i-1} = 2N(1 - \frac{1}{k}) \tag{1.22}$$

Thus the average time to the most recent common ancestor of a haploid population is in order of $2N$ generations in a Wright-Fisher model, and $4N$ for a diploid population.

Note that the t_i are random variables and that the whole process is highly stochastic. Tavaré (1984) computed the whole distribution of the Kingman coalescent TMRCA, giving:

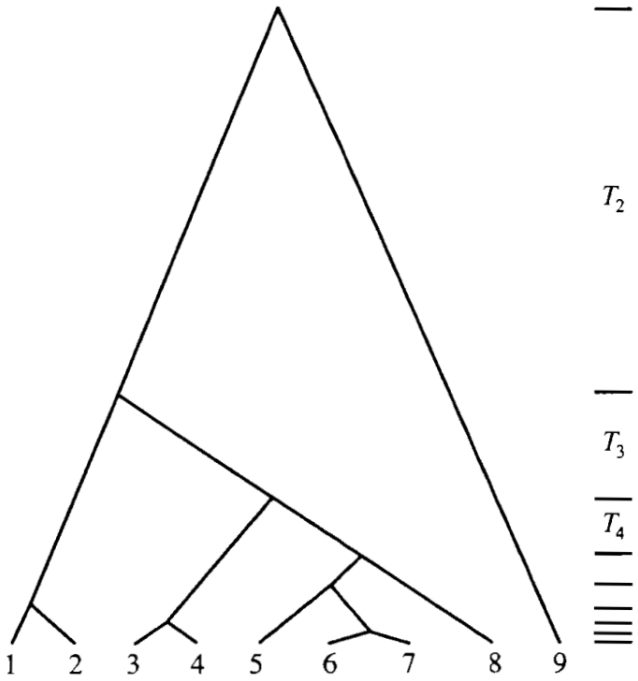

Figure 1.3: **Example of a Kingman coalescent tree of a sample of 9 individuals. Extracted from Wakeley (2009).** The time interval t_k during which there is k individuals is represented on the right-hand side of the tree.

$$f_{TMRCA}(t) = \sum_{i=2}^{n} \frac{(2i-1)(-1)^i(n(n-1)...(n-i+1))}{n(n+1)...(n+i-1)} \binom{i}{2} \mathrm{e}^{-\binom{i}{2}t} \tag{1.23}$$

The distribution of T_{MRCA} is provided in Figure 1.4. It shows that the distribution is skewed towards smaller T_{MRCA} values. While the approximated expected value is quite robust to variations in population size, the realized value is most of the time smaller. Note also the broadness of the distribution that illustrates the stochasticity of random genetic drift.

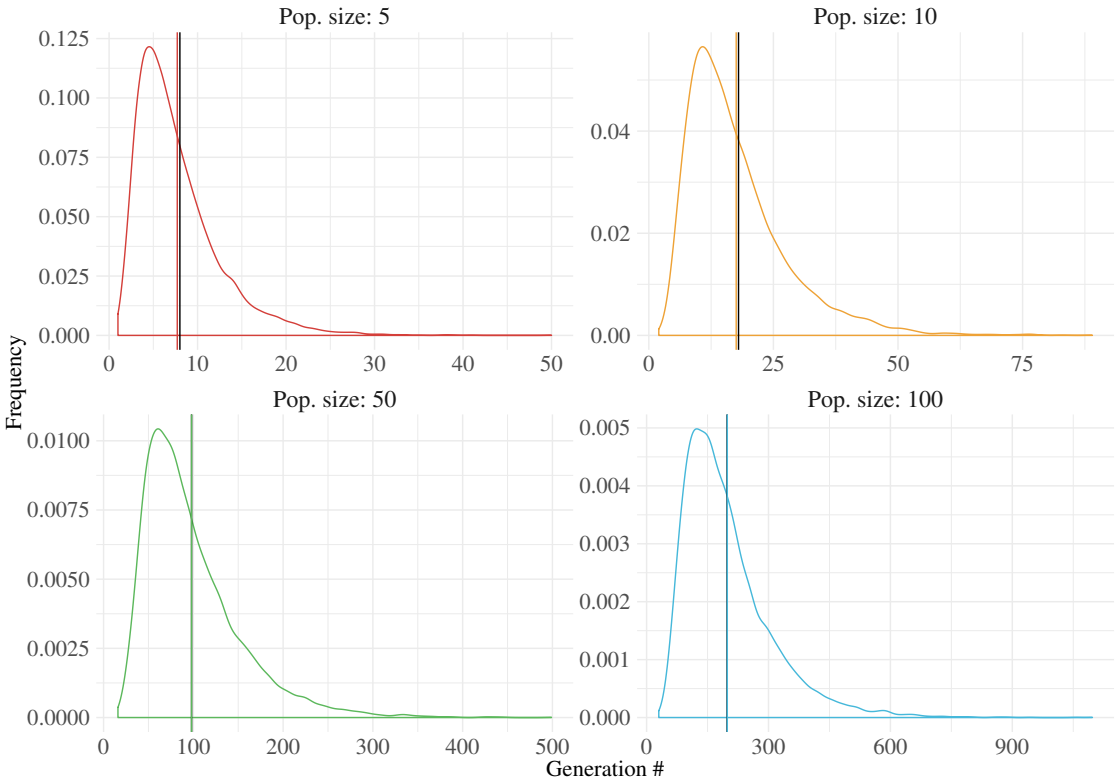

Figure 1.4: T_{MRCA} **distribution in** N **generation, for** N **and k in** $\{5, 10, 50, 100\}$**.** Each distribution has been produced by simulating the random coalescence of $k = N$ haploid individuals. Each black vertical line corresponds to the expected value: $\mathbb{E}(T_{MRCA}) = 2N(1 - \frac{1}{k})$ and the colored vertical lines correspond to the simulated mean. The expected value, neglecting multiple merging is quite robust to the population size.

1.2.2 Mutation

In the previous paragraph, we have seen that genetic drift alone leads to the inescapable loss of genetic diversity due to allele fixation or loss. However, it is particularly obvious from phenotypic or molecular observations that a certain amount of diversity is observable. Mutations are modifications of the genetic information that occur during DNA replication and result in the occurrence of a new allele. Evolutionary relevant mutations are the ones transmitted to offspring, and therefore population genetics focuses more on mutations in germinal cells. (N.B. Population genetics models are now used to study the fate of other types of mutations, like somatic mutations occurring in different types of cancer, considering cancerous tissues as population of daughter cells).

We can distinguish several types of causes for mutations: spontaneous mutations (molecular decay, like depurination or deamination), mutations during repair of DNA damage (like error-prone translesion synthesis involved in the repair of apurinic site, or through non-homologous end joining repair process), induced mutations caused by mutagens (like DNA intercalating agents, oxidative chemicals, UV, ionizing radiations...), transposable elements movements, etc...

These different causes for mutation occurence lead to several classes of DNA modifications:

- Substitutions of one nucleotide by another. The change of a purine (A or G) resp. pyrimidine (C or T) to another purine resp. pyrimidine is called transition. The change from a purine to a pyrimidine (and vice versa) is called transversion. Transitions and transversions occur at different rates, with the former being more frequent than the later.

- Insertions of one to many nucleotides (several kb). Insertions are often caused by transposable elements, or errors during replication of repeating elements. Insertions in the coding region of a gene may alter splicing of the mRNA (splice site mutation), or cause a shift in the reading frame.

Insertions often significantly alter the gene product.

- Deletions of one to many nucleotides. Deletions often cause a shift in the reading frame, altering the gene product. Large deletions of DNA can lead to the loss of one to several genes.
- DNA duplication, like gene duplication.
- Chromosomal translocations: the interchange of large piece of DNA between non-homologous chromosomes.
- Chromosomal inversions: reversing the orientation of a chromosomal segment.
- Etc.

In any case, the appearance of a new allele results from a single event occurring by chance. The initial frequency of the mutation is $p_0 = 1/N$ in haploids, and $p_0 = 1/2N$ in diploids. The fate of a new mutation is determined by the other evolutionary forces.

1.2.2.1 Independence of mutation occurrence from selection

The demonstration that genetic mutations arise randomly in population, independently of other evolutionary forces was brought by the famous Luria–Delbrück experiment (Luria and Delbrück, 1943), also called the Fluctuation Test. In this experiment, a small number of cells from a *E. Coli* strain sensitive to a phage were used to inoculate a culture in a non selective liquid medium, in the absence of the phage. After the stationary phase was reached, samples were plated on a rich selective medium in presence of the phage. After some time, the number of colony forming units was counted on each plate. Because the medium was selective, only resistant cells could form a colony (Fig. 1.5)

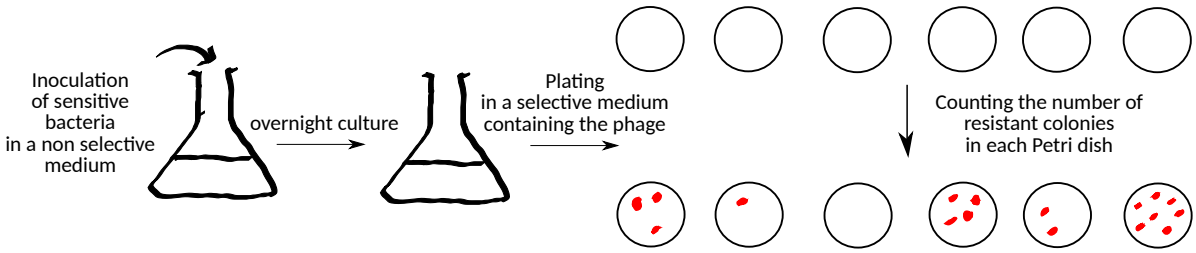

Figure 1.5: **The Luria-Delbrück experiment**

Let X_i be the number of resistant colonies on plate i. There were two hypotheses (Fig. 1.6):

Were mutations induced by selection? In this case, mutations could only appear in the plated cells, and each cell would have the same tiny probability μ of getting a mutation. Hence, X_i would result from the drawing in a binomial distribution. Furthermore, because μ is expected to be very small compared to the number of cells plated, X_i can be approximated as following a Poisson distribution with its variance being equal to its mean.

Were mutations spontaneous, and independent of the selective agent? In that case, they could appear at any time from the beginning of the experiment. In particular, a mutation appearing in the pre-culture and having the same growth rate as the wild-type cells may reach a significant census size at the end of the overnight culture. Plating here consists on sampling randomly n cells, among which a fraction p_0 can carry the resistance gene. Because of drift, the authors expected a variance between plates much larger than the mean.

This experiment clearly demonstrated the independence of mutation occurrence from the selective agent, as the experimental results experiment clearly fitted the hypothesis that mutations occur at random, prior to the plating with the selective agent. In this experiment, their fate was determined by genetic drift and selection.

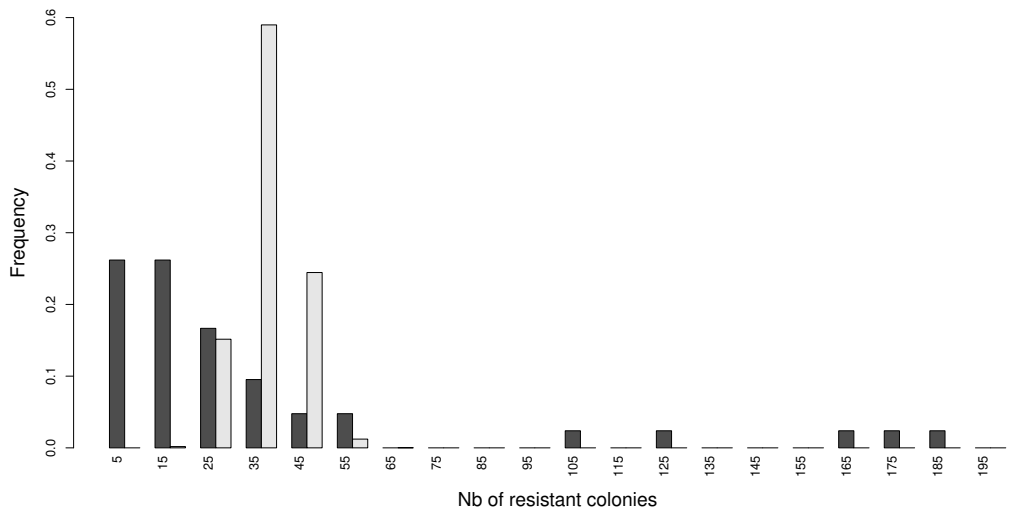

Figure 1.6: **Distribution of the number of resistant colonies in the Luria-Delbrück experiment.** The figures from the paper are in black. The gray distribution corresponds to a Poisson distribution with mean equal to the observed mean.

1.2.2.2 Population genetic model for neutral mutations

Any mutation can be modeled using a di-allelic model, in which one allele is mutated and the others not. From this vantage point, the fate of an allele can be described without reference to its specific molecular basis. Nevertheless, when several mutations occur at the same locus or different types of mutations are considered, adjustments of the parameters (*e.g.* mutation rate) should be considered, as discussed below.

The simple di-allelic model supposes A the randomly mutated version of allele a. Consider a probability of mutation, or mutation rate, μ. What is the expected allele frequency of A at time $t+1$, p_{t+1}, knowing p_t ?

$$\mathbb{E}(p_{t+1}|p_t) = p_t + \mu(1-p_t) \tag{1.24}$$

Therefore, the expected change of allelic frequency of A, at each generation is $\mathbb{E}(p_{t+1} - p_t) = \mu(1-p_t)$ if no other forces are involved. Tacking into account genetic drift and the recurrent appearance of A alleles by mutation, all individuals in the population will eventually be of type A.

If mutation can occur in both directions, so that $\mu_{A\to a}$ corresponds to the probability that A mutates into a, and $\mu_{a\to A}$, the opposite:

$$\mathbb{E}(p_{t+1}|p_t) = (1-\mu_{A\to a})p_t + \mu_{a\to A}(1-p_t) \tag{1.25}$$

In the absence of other forces, an equilibrium will be reached when $\mathbb{E}(p_{t+1}|p_t) = p_t = p^*$ and it follows:

$$p^* = \frac{\mu_{a\to A}}{\mu_{A\to a} + \mu_{a\to A}} \tag{1.26}$$

A broad application of this simple population genetics model is the inference of phylogenetic trees from molecular data, taking into account different mutation rates for each possible DNA base substitution (Kimura, 1980). Building on the transition matrix, Felsenstein (1981) proposed a maximum likelihood method to infer evolutionary trees from DNA sequence data that takes into account the possibility of different rates of evolution in different lineages.

1.2.2.3 Mutation rate estimations

The straightforward method to estimate mutation rate is to perform mutation accumulation (MA) experiments. In this type of experiment, a single inbred and highly homozygous line is typically replicated and maintained for many generations at small population size (through brother and sister mating, selfing, etc.), in an environment designed to minimize the effects of natural selection. As mutations are accumulating, variance between lines increases. The phenotypic changes across generations is used to indirectly estimate the mutation rate for that organism. If one wish to estimate the mutation rate affecting a simple phenotype due to monogenic resistance, one could confront the derived lines to the selective agent, as in the Fluctuation Test, to estimate the proportion of resistant individuals. Knowing the breadth of the mutational target, one can further estimate the mutation rate per base pair (by dividing the mutation rate by the number of bases of the mutational target). The actual advances in sequencing techniques combined with mutation accumulation experiments improve genome-wide estimates of mutation rate: as we have seen in the drift section, neutral mutations become fixed at the mutation rate μ; and therefore, by looking at the number of fixed synonymous (neutral) mutations, one can directly estimate the neutral mutation rate (Wielgoss et al., 2011). For example, the famous Lenski's *E. coli* long term-evolution experiment that started in 1988 and reached over 60 000 generations in 2016 with 12 populations grown on different media (Good et al., 2017), allowed Wielgoss et al. (2011) to produce a reliable estimate of 8.9×10^{-11} point mutation per base-pair per generation. However some uncertainties still remain and variance in the mutation rate within species can be found in the literature, among different bacterial strains for example. Note that the mutation rate varies greatly along the genome, influenced by factors such as the recombination rate, and the presence of repeats (short sequence repeat, transposable elements) (Drake et al., 1998).

Mutation rate can also be inferred from phylogenetic data thanks to the common hypothesis of constant accumulation of neutral mutation (clock-like) in a genome: the **molecular clock** hypothesis. The phylogenetic method uses a known phylogenetic tree of closely related taxa and their approximate TMRCA. If we count neutral mutations between two taxa and assume that these mutations have been accumulating at a constant rate since the TMRCA, we can estimate the mutation rate. Indeed, as we have seen previously, the rate of fixation of neutral mutation under genetic drift is equal to the mutation rate. For further discussion see for example Scally and Durbin (2012).

Table 1.1, compile mutation rates estimates, for several model organism and through different protocols.

Species	Tissue	Cell divisions per generation	Mutation rate per generation	Mutation rate per cell division	Source
Escherichia coli		1	0.26×10^{-9}	0.26×10^{-9}	Lynch (2010)
		1	0.089×10^{-9}	0.089×10^{-9}	Wielgoss et al. (2011)
		1	0.50×10^{-9}	0.50×10^{-9}	Lee and Palsson (2010)
		1	0.15×10^{-9}	0.15×10^{-9}	Kishimoto et al. (2010)
Saccharomyces cerevisiae		1	0.33×10^{-9}	0.33×10^{-9}	Lynch (2010)
Caenorhabditis elegans	Germline	9	5.60×10^{-9}	0.62×10^{-9}	Lynch (2010)
Drosophila melanogaster	Germline	36	4.65×10^{-9}	0.13×10^{-9}	Lynch (2010)
Mus musculus	Germline	39	38.00×10^{-9}	0.97×10^{-9}	Lynch (2010)
Homo sapiens	Germline	216	12.85×10^{-9}	0.06×10^{-9}	Lynch (2010)
Arabidopsis thaliana	Germline	40	6.50×10^{-9}	0.16×10^{-9}	Lynch (2010)
Zea mays ssp. mays			$29 - 33 \times 10^{-9}$		Clark et al. (2005)

Table 1.1: **Diversity of mutation rate (substitution) across living organisms.** Note that in eukaryotes, mutations that are transmitted are the ones occurring in the germinal cells. Because cellular divisions occur in the germline, the mutation rate per generation is higher than the mutation rate per cell.

1.2.2.4 Mutation rate evolution

From table 1.1, the mutation rate seems to stay "low" throughout the tree of life, *i.e.* a 10^{-9} mutation probability seems consistent with the first role of the DNA, that is to convey reliably the genetic information. However, such small probabilities are difficult to apprehend, for example, what is the evolutionary impact of a change in mutation rate? and how can we explain differences between organisms? *E.g.* Prokaryotes also seem to have a lower mutation rates per generations than eukaryotes.

To answer those questions, we can correlate these differences with other parameters, such as the genome size. Drake (1991) concluded that the mutation rate per base pair per generation scales inversely with the genome size (G) in DNA-based microbes, which have been confirmed for double-stranded DNA viruses and other prokaryotes (Lynch, 2010). However in contrast to prokaryotes, eukaryotic mutation rates scale positively with genome size, with vertebrates rates being nearly 100 times higher than that of prokaryotes, rates for unicellular eukaryotes, invertebrates, and land plants being intermediate.

The evolution of mutation rate results from several interacting forces. High mutation rates increase the number of deleterious mutations and tend to be selected against. But a high mutation rate also produces more advantageous mutations. Low mutation rates may also result in metabolic cost and slow replication rate due to the involvement of more accurate repair systems. The idea of a trade-off was explored (Kimura, 1967), however the expected mutation rate should be lower than what is observed. One possible explanation is that the power of genetic drift (in order of $1/N_e$, the effective population size, a concept which will be explained later on) ultimately constrains what natural selection can accomplish with any trait. Once the incremental effect of reduced mutation rate is smaller than the power of drift, the mutation rate reaches a lower bound. A better understanding of the evolution of the mutation rate would require a fine comprehension of the interplay between drift and selection (Lynch, 2010; Lynch et al., 2016).

1.2.2.5 Mutational effect on fitness and its distribution

Point mutations are often characterized as non-coding (if they occur in non coding region like introns) or coding (e.g. in exons). Coding mutations are then decomposed in non-synonymous or synonymous, if the change of nucleotide change (resp. do not change) the coded amino-acids. Note that these DNA modifications can lead to several effects on function, like loss-of-function (e.g. the insertion of a STOP codon), gain-of-function, function modification, changes in levels of expression or dominance.

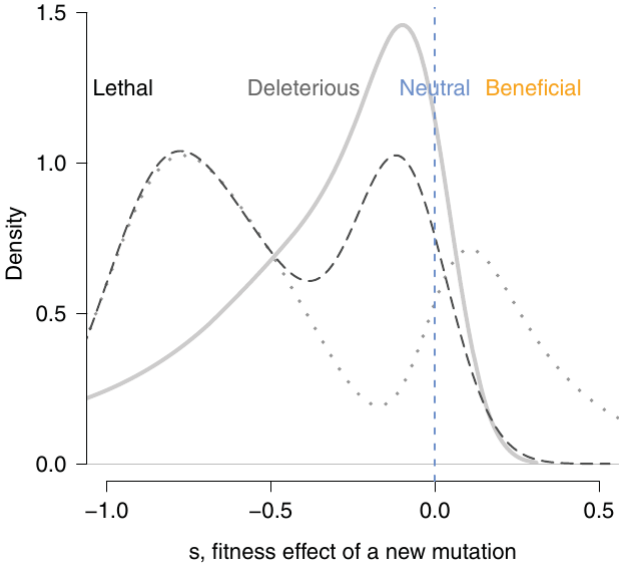

Figure 1.7: **Hypothetical DFEs.** The distribution of fitness effects comprise a continuum of effects from lethal, strongly or mildly deleterious, to beneficial. Three distributions are pictured here that comprise different amounts of deleterious and beneficial mutations: the dashed distribution where most mutations are detrimental with two modes, one for lethal and the other for deleterious mutations; the dotted distribution, also bimodal, which allows for substantial proportion of beneficial mutations; the continuous distribution (solid gray) is predicted by a Fisherian fitness landscape (Figure from Bataillon and Bailey (2014)).

We can distinguish three classes of mutations according to their effect on fitness: (i) mutations that confer a reproductive disadvantage to the bearer and are selected against, called lethal or deleterious mutations depending on their strength; (ii) mutations that have no fitness effects and evolve under drift, called neutral mutations; and (iii) mutations that confer a reproductive advantage to the bearer and are selected for, called advantageous mutations. However, these three categories conceal a continuous mutational fitness effect distribution. The distribution of these fitness effect (DFE), is of particular importance to geneticists Bataillon and Bailey (2014), as it caries information needed to understand the nature of quantitative genetic variation, and the dynamics of *de novo* mutations in adaptation, and to predict the consequences of maintaining animals or plants at low population, etc... Figure 1.7 illustrates three hypothetical distributions of fitness effects. There are two main difficulties in inferring the DFE. First, we expect that, within a well-adapted population, most mutations with detectable fitness effects will be deleterious. Second, if beneficial mutations are rare, they consist in extreme events drawn in the right-tail of the distribution. The Extreme Value Theory (EVT) tells us that draws from tails of virtually any distribution converge to the same distribution called the generalized Pareto distribution (Bataillon and Bailey, 2014; Beisel et al., 2007), so that different distributions can hardly be compared.

Despite those difficulties, two approaches, mutation accumulation and mutagenesis experiments on one hand, and the analysis of DNA sequence data, on the other hand, have been used to decipher the DFE, as reviewed in Eyre-Walker and Keightley (2007). The main idea in the analysis of polymorphism and divergence data remains rather close to one of the first attempts developed by (McDonald and Kreitman, 1991) to measure the amount of selection experienced by genes, the McDonald-Kreitman test. This test computes the ratio d_n/d_s of non-synonymous variation (supposed to experience selection) over the synonymous variation (supposed to evolve neutrally). It then compares the amount of variation between

species (d_n/d_s) to the amount of variation within the sampled species (π_n/π_s). Some authors, like Smith and Eyre-Walker (2002), then built on this test to infer α the proportion of advantageous substitutions, and the proportion of strongly deleterious substitutions experiencing purifying selection. Then using the theory of Poisson random field, Huerta-Sanchez et al. (2008) (see Sethupathy and Hannenhalli (2008) for a theoretical tutorial), proposes an estimate of the DFE. The Site Frequency spectrum (SFS, a concept developed later) is used to better integrate patterns of nucleotide variation within the focal species, while variation between the focal species and the outgroup is measured by counting the number of fixed mutations between them. (The SFS can be linked to the DFE and demographical parameters like the effective population size; selection is assumed to be weak ($s << 1$) and the DFE to be constant in time and the same in both the ingroup and outgroup). Selection and demography are again disentangled by separating the sites into neutrally evolving and selected sites classes. The DFE is then inferred by contrasting the SFS counts for the neutral and selected sites, by assuming homogeneity in demography and other forces, between the two classes. One can note the method developed by Tataru et al. (2017) that uses a hierarchical probabilistic framework that extends previous methods to infer DFE and α from polymorphism data alone, without taking into account any information from an outgroup. They used a mixture between gamma and exponential distributions, which model deleterious and beneficial mutations, respectively, according to probability parameters.

We need to keep in mind the difficulties of estimating the DFE. First the effect of mutations are difficult to disentangle from demography, as neutral mutations depend on their ability to overcome drift. The concept of effectively neutral mutations has emerged in that sense. Secondly, the advantageous or deleterious qualification of a mutation depends on the adaptive state of the population. In conditioning the effect of mutations on a fitness function of a population, we can distort the absolute DFE of mutations, as a strongly advantageous mutation might overshoot the optimal value, hence become deleterious if the individual is already well adapted (see Fisher Geometric model hereinafter). Finally, the proportion of mutations that are advantageous, effectively neutral and deleterious may vary between species, and differ between coding and non-coding regions.

However, as reported by Eyre-Walker and Keightley (2007), we can draw the following conclusions:

- Advantageous mutations are rare, and those that are strongly selected are exponentially distributed, as suggested by Orr and Gillespie from the Extreme Value Theory (for example see: Joyce et al. (2008)). This property remains nevertheless difficult to test because of a lack of power emanating from the rarity of this mutational class.
- The DFE of deleterious mutations is complex and multi-modal.
- It is unlikely that mutations are truly neutral, they however have vanishingly small effects. But effectively neutral mutations are easier to identify through the joint estimation of demographical parameters.

1.2.3 Recombination

We focused previously on one locus mainly, and didn't account for the chromosomal arrangement of loci that are not transmitted independently, so that linked loci segregate together during meiosis. In order to take into account this non independent transmission, we can first define, the recombination rate c between two loci as the proportion of recombinant gametes after meiosis. c varies between 0 (same locus) and 1/2 (independent loci). By shuffling the combination of paternal and maternal alleles along the chromosome, recombination results in new gametic combinations in a population. As shown below, it also changes gamete frequencies. For those two reasons, we can consider recombination as an additional evolutionary force.

1.2.3.1 Linkage Disequilibrium

Definition and measurement Let's first consider a simple case, a diploid population constituted of N individuals with two loci: A and B, with respectively two alleles A,a and B,b. The four combinations are called haplotypes (AB, Ab, aB, ab). Let's count the number of individuals presenting each haplotype:

n_{AB},n_{Ab},n_{aB} and n_{ab}, and their respective frequencies:f_{AB},f_{Ab},f_{aB} and f_{ab}. We can retrieve each allelic frequency by summing the corresponding haplotype frequencies, such as $f_A = f_{AB} + f_{Ab}$.

If alleles at each locus were independent, the probability of any haplotype would be the product of its corresponding allele frequencies: $f_{AB} = f_A \times f_B$. However, if the two alleles A and B are not independently inherited, one can measure the departure from this equilibrium, by introducing **D, the linkage disequilibrium coefficient** (Bennett, 1952), such as:

$$D_{AB} = f_{AB} - f_A \times f_B \tag{1.27}$$
$$D_{Ab} = f_{Ab} - f_A \times f_b \tag{1.28}$$
$$D_{aB} = f_{aB} - f_a \times f_B \tag{1.29}$$
$$D_{ab} = f_{ab} - f_a \times f_b \tag{1.30}$$

Note that when $D_{AB} = 0$, the loci are said to be in linkage equilibrium. Using the sum property of haplotype frequencies, one can easily show that:

$$D_{Ab} = D_{aB} \text{ and } D_{AB} = D_{ab} \text{ and } D_{AB} = -D_{Ab} \tag{1.31}$$

Therefore, one can consider only one value, that we call D, without any subscript and whose sign is dependent on which allele is first considered like $D = D_{AB} = D_{ab}$. How can we interpret the value of D ? We can see that the value of D is constrained by the allele frequencies:

$$f_{AB} = f_A \times f_B + D \tag{1.32}$$
$$f_{Ab} = f_A \times f_b - D \tag{1.33}$$
$$f_{aB} = f_a \times f_B - D \tag{1.34}$$
$$f_{ab} = f_a \times f_b + D \tag{1.35}$$

As the haplotype frequencies cannot be negative, we have:

$$D > 0 \Rightarrow D \leq \min(f_A \times f_b, f_a \times f_B) \tag{1.36}$$
$$D < 0 \Rightarrow -D \leq \min(f_A \times f_B, f_a \times f_b) \tag{1.37}$$

Therefore, following Franklin and Lewontin (1970), D can be normalized by its maximum value, and we obtain:

$$\left.\begin{array}{l} D' = \frac{D}{\min(f_A \times f_b, f_a \times f_B)} \text{ if } D > 0 \\ D' = \frac{-D}{\min(f_A \times f_B, f_a \times f_b)} \text{ if } D < 0 \end{array}\right\} \Rightarrow 0 \leq D' \leq 1 \tag{1.38}$$

If $D' = 1$, $|D|$ is maximal, and therefore, at least one haplotype is missing.

One other traditional way to measure linkage disequilibrium consists on using a correlation coefficient whose significance can be easily tested using a χ^2 test (Hill, 1975) . *E.g.*, let's consider a couple of random variables (X_i, X_j) so that $X_i = 1$ if an haplotype is composed of an allele A and $X_i = 0$ if it includes an a and $X_j = 1$ if the haplotype presents an allele B and $X_j = O$ if it presents a b.

$$\mathbb{E}(X_i) = 1 \times \Pr(X_i = 1) + 0 \times \Pr(X_i = 0) = f_A \tag{1.39}$$
$$\mathbb{E}(X_j) = 1 \times \Pr(X_j = 1) + 0 \times \Pr(X_j = 0) = f_B \tag{1.40}$$
$$\mathbb{E}(X_i X_j) = 1 \times 1 \times f_{AB} + 1 \times 0 \times f_{Ab} + 0 \times 1 \times f_{aB} + 0 \times 0 \times f_{ab} = f_{AB} \tag{1.41}$$

$$\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 \tag{1.42}$$
$$= 1 \times \Pr(X_i = 1) + 0 \times \Pr(X_i = 0) - f_A^2 \tag{1.43}$$
$$= f_A - f_A^2 = f_A(1 - f_A) = f_A f_a \tag{1.44}$$
$$\mathbb{V}(X_j) = f_B f_b \tag{1.45}$$

Then:

$$r^2(X_i, X_j) = \frac{(\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\,\mathbb{E}(X_j))^2}{\mathbb{V}(X_i)\,\mathbb{V}(X_j)} = \frac{(f_{AB} - f_A f_B)^2}{f_A f_a f_B f_b} \tag{1.46}$$

$$r^2(X_i, X_j) = \frac{D^2}{f_A f_a f_B f_b} \tag{1.47}$$

- Because D depends on the allele frequencies, we expect values of D to vary a lot over many pairs of markers, even when sites are in complete LD.
- D' has the interesting property to always be equal to 1.0 if two SNPs are in complete LD. Furthermore, we only need three gametic types without the need to know the recombination rate to explain patterns of variation. However, because of the normalization by the product of marginal allele frequencies, the sampling properties of D' are difficult to study (Gaut and Long, 2003).
- We will note that these measures of LD are derived only for 2 loci, with an increasing complexity with more markers and more alleles per site. One way traditional way to overcome this limitation is through the summary of pairwise measures of LD in matrix form, or to plot values of r^2 against a physical map. New interesting definitions have been proposed, such as the paradoxical definition of linkage disequilibrium at one locus, that allows to see LD as the probabilistic independence of certain events (Gorelick and Laubichler, 2004).

Linkage Disequilibrium evolution : recombination If we are interested in the evolution of LD through generations, we can for example consider a diploid, panmictic, population of infinite size without selection or mutation, and compute the frequency of f_{AB}, after one generation of panmixia.

First we need:

$$f_{AB}f_{ab} - f_{Ab}f_{aB} = [f_A \times f_B + D][f_a \times f_b + D] - [f_A \times f_b - D][f_a \times f_B - D] \tag{1.48}$$

$$= [f_A f_B + f_a + f_b + f_A f_b + f_a f_B]D \tag{1.49}$$

$$= D \tag{1.50}$$

Let's compute the frequency of f_{AB}, after one generation of panmixia. After meiosis, AB gametes can be produced without recombination by any genotype carrying a AB chromosome: (AB/AB, AB/Ab, AB/ab, AB/ab). But AB gametes can also be created by recombination from the following genotype : (Ab/aB, aB/Ab, AB/AB, AB/Ab, AB/aB). If we count all possible crosses between haplotypes that lead to AB, we get:

$$f_{AB}(t+1) = f_{AB}(t)^2 + f_{AB}(t)f_{Ab}(t) + f_{AB}(t)f_{aB}(t) + f_{AB}(t)f_{ab}(t) + c \times f_{Ab}(t)f_{aB}(t) - c \times f_{AB}(t)f_{ab}(t) \tag{1.51}$$

$$= f_{AB}(t) - c[f_{AB}(t)f_{ab}(t) - f_{Ab}(t)f_{aB}(t)] \tag{1.52}$$

Recalling that $f_{AB}(t) = f_A f_B + D(t)$, and that in the absence of selection, drift, mutation, we have $f_A(t+1) = f_A(t)$, we finally obtain:

$$f_A f_B + D(t+1) = f_A f_B + D(t) - cD(t) \tag{1.53}$$

Which leads to:

$$D(t+1) = (1-c)D(t) \tag{1.54}$$

By induction on t, we have:

$$D(t) = (1-c)^t D(0) \tag{1.55}$$

Therefore, linkage disequilibrium between two loci decreases on average at a rate $(1-c)$ per generation. However, if one generation of panmixia is enough to create Hardy-Weinberg equilibrium, a much longer

time is required to reach linkage equilibrium. The decrease in LD after $1/c$ generations tends towards 37% as c tends towards zero. This means that, whatever the evolutionary force creating linkage disequilibrium between two loci, two closely linked loci may stay in strong LD for a period of time inversely proportional to their recombination rate.

> In panmictic population, the amount of linkage disequilibrium between pairs of loci is expected to be inversely proportional to their recombination rate. Because of recombination and Mendelian segregation, independent loci are expected to be in linkage equilibrium.

This consideration had important consequences in population and quantitative genetics. It is at the basis of genetic maps construction and of Genome Wide Association Studies (GWAS).

Linkage disequilibrium : drift and mutation Both drift and mutation are evolutionary forces that create linkage disequilibrium in populations.

As seen before, a mutation occurs during reproduction and results in the modification of a DNA sequence in the progenies of a single individual. Let's consider a polymorphic locus with two alleles A (in frequency $p(t)$) and a in a diploid population of size N. Suppose that a mutation occurs at a neighboring locus that changes the monomorphic allele b into a new allele B in an A individual. At the next generation, there will be three haplotypes having the following frequencies:

haplotype	AB	Ab	aB	ab
frequency	$1/N \cdot p(t)$	$(1-1/N)p(t)$	$1-p(t)$	0

and the linkage disequilibrium will be $D(t+1) = p(t)/N$, with r^2 close to one: the fate of the A allele will be linked to the fate of the new B allele. If B is a favourable mutation, haplotype AB will increase in frequency and make the frequency of allele A also increase in the population.

Random genetic drift is also a powerful mechanism that may create linkage disequilibrium. Indeed, in finite populations, the frequency of haplotype AB in the next generation is the result of a Binomial$(N, f_{AB}(t))$ sampling. Haplotype frequencies are going to fluctuate as in a brownian motion. For neutral loci, Hill and Robertson (1968) linked r^2 to the recombination rate through the widely used parameter $4N_e c$ in diploids population, so that:

$$\mathbb{E}(r^2) = \frac{1}{1+4N_e c} \tag{1.56}$$

where c is the recombination rate in morgans between the two markers and N_e is the effective population size. This equation emphasizes that LD is a population properties and should be interpreted as such.

In particular, the fate of a new mutation that appears in a given haplotype context will determine the fate of the alleles at all neighbouring loci. Both beneficial and detrimental mutations will tend to decrease the genetic polymorphism at surrounding loci by accelerating their fixation or loss. However, the two dynamics are different. The fixation of rare beneficial mutations will lead to a local decay of polymorphism called *selective sweep* (Maynard Smith and Haigh, 1974). Because slightly detrimental mutation may occur at a higher frequency at the genome scale, the loss of detrimental mutations will lead to a general decay of neutral polymorphism concentrated in regions with low recombination. This phenomenon is called *background selection* (Charlesworth et al., 1993). Selective sweeps will be seen more in detail in the next chapter.

1.2.3.2 Recombination and crossing-overs

Genetic recombination is the genetic reshuffling of combination of several locus that leads to production of offspring with combinations that differ from those found in either parent. In eukaryotes, genetic recombination occurs during meiosis, through a process called chromosomal crossover, or crossing over. It corresponds to the exchange of genetic material between two homologous chromosomes non-sister chromatids that results in recombinant chromosomes during sexual reproduction, occasionally producing new chimeric alleles.

Crossing-overs During the pachytene stage of prophase I, this process involves the formation of a synaptonemal complex, which is a protein structure that forms between homologous chromosomes (two pairs of sister chromatids) to mediate chromosome pairing. Crossover usually occurs when matching regions on matching chromosomes break and then reconnect to the other chromosome, thanks to a structure called "Holliday junction". The resolution of this double-strand break initiate the recombinational repair process. The repair of the gap can lead to two different mechanisms: crossover repair or non-crossover repair of the flanking regions. Crossover repair occurs by the Double Holliday Junction (DHJ) model while non-crossover repair occurs primarily by the Synthesis Dependent Strand Annealing (SDSA) model, leading to gene conversion. See figure 1.8.

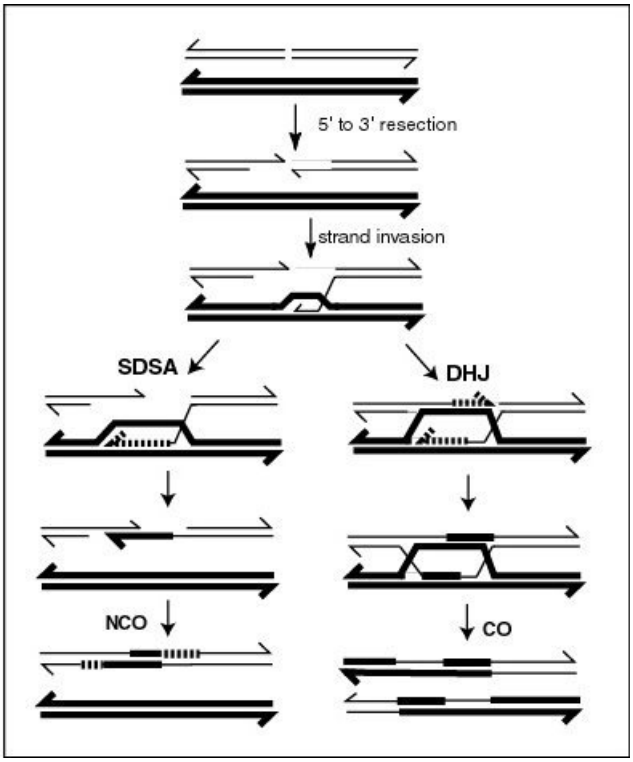

Figure 1.8: Schematic view of the DNA repair process leading to crossover or non-crossover mechanisms. Crossing over leads to recombination and non-crossing over to gene conversion. From Bernstein et al. (2011).

Because recombination can occur with small probability at any location along chromosome, the frequency of recombination between two locations depends on the distance between them. Recombination points can be simulated by considering linear chromosomes and drawing successive recombination points in an exponential law of parameter λ, so that $1/\lambda$ is the average number of recombination event per chromosome Falque et al. (2007). Note that this model supposes no interference.

Recombination pattern along chromosomes and the crossing-over interference Among other molecular particularities of recombination Zickler and Kleckner (2015) reviewed the distribution pattern of crossing over along chromosomes. They discuss the fact that both double-stranded breaks and consequent crossovers tend to be evenly spaced along the chromosomes by mechanisms that remain to be discovered. Those mechanisms seem to prevent two crossovers to fall nearby. This feature was first discovered in Drosophila and called “crossover interference” (Muller, 1916; Berchowitz and Copenhaver, 2010). *I.e.* if a crossover happened at one loci, there is a reduced probability that another crossover will occur nearby, but this probability increases again as distance increase. This implies the existence of communication along the chromosomes, the basis for which is still incompletely known. However, we can conservatively observe that regular homologous segregation in meiosis requires at least one crossover (chiasma): a first “obligatory crossover” (Jones and Franklin, 2006).

Several models have been proposed, 3 major are based on interference mechanisms:

- King and Mortimer (1990): a model based on the kinetics of crossover designation and signal spreading.
- Lande and Stahl (1993): a model based on a mechanism that begins at one end of a chromosome and "count" double stranded breaks interactions, with crossover-designation occurring after a specific (nearly) fixed number of precursors.
- Kleckner et al. (2004); Zhang et al. (2014): a model based on the idea that communication for crossover interference might occur via redistribution of mechanical stress: the "beam film model".

Another way of modeling crossover patterning is to try to match the observed pattern to any particular mathematical distribution. Considerable attention has been given to modeling by the Γ distribution, (often used to analyze crossover patterns), as the "signal" seems to decrease exponentially with distance away from the nucleating site, and that the sum of i.i.d. exponential distribution is a special case of Γ distribution (McPeek and Speed, 1995; Falque et al., 2007). This approach only describes the final outcome of the process, without regard to any other features.

1.2.4 Selection

In this paragraph, we will try to understand how selection, *i.e.* differences on survival and reproduction between individual, changes the allele frequencies in a population and what are its impact on the phenotypic values.

1.2.4.1 Haploid model

Again, to build our intuition, we will start by considering selection at one locus, with two possible alleles A and a, in a haploid population with discrete non-overlapping generations. In this first approach, we will not take into consideration the stochasticity in the number of offspring of an individual, we will only consider a purely deterministic model.

Let $N_A(t=0)$ be the number of individuals of type A at time $t=0$ and $N_a(t=0)$ be the number of individuals of type a. This translates into the following definition of allele frequencies:

$$f_A(t=0) = \frac{N_A(t=0)}{N_A(t=0)+N_a(t=0)} \text{ and } f_a(t=0) = \frac{N_a(t=0)}{N_A(t=0)+N_a(t=0)} \tag{1.57}$$

Let w_A the number of offspring of individuals carrying the allele A individual and w_a for individuals carrying a. Here the $w's$ are also called absolute fitness, and they depend on the probability that an individual survives to reproduction and the probability that it reproduces, both of which are under the influence of environmental conditions, a combination of biotic and abiotic factors. However to decipher the factors leading to differences in absolute fitness won't be relevant in our case. Furthermore, what really matters are the differences in $w's$ values between the two genotypes. To see this, let's try to compute the change in allele frequency after one generation:

$$f_A(t=1) = \frac{N_A(t=1)}{N_A(t=1)+N_a(t=1)} \tag{1.58}$$

$$= \frac{w_A N_A(t=0)}{w_A N_A(t=0)+w_a N_a(t=0)} \tag{1.59}$$

Dividing numerator and denominator by $N_A(t=0)+N_a(t=0)$ gives:

$$\tag{1.60}$$

$$= \frac{w_A f_A(t=0)}{w_A f_A(t=0)+w_a f_a(t=0)} \tag{1.61}$$

Dividing numerator and denominator by w_A gives:

$$\tag{1.62}$$

$$= \frac{f_A(t=0)}{f_A(t=0)+\frac{w_a}{w_A} f_a(t=0)} \tag{1.63}$$

We can see from this last line, that the evolution of allele frequencies, only depends on the ratio between absolute fitness. A common way of seeing this problem consists in defining a **selection coefficient** s so that: $1 - s = \frac{w_a}{w_A}$

$$f_A(t=1) = \frac{f_A(t=0)}{f_A(t=0) + (1-s)f_a(t=0)} \tag{1.64}$$

If we suppose that the differences in reproductive success persists through time, we can show that:

$$N_A(t) = w_A^t N_A(t=0) \tag{1.65}$$

Therefore, by induction on t, we can write:

$$f_A(t) = \frac{f_A(t=0)}{f_A(t=0) + (1-s)^t f_a(t=0)} \tag{1.66}$$

Figure 1.9 plot the evolution of the an allele A with varying selection coefficients from $f_A = 0.05$ through generations.

From this figure and the limit when t goes to infinity in the last equation, we can see that if an allele confers even a small increase in the number of offspring to its bearer, and no other forces acts upon it, it will eventually goes to fixation.

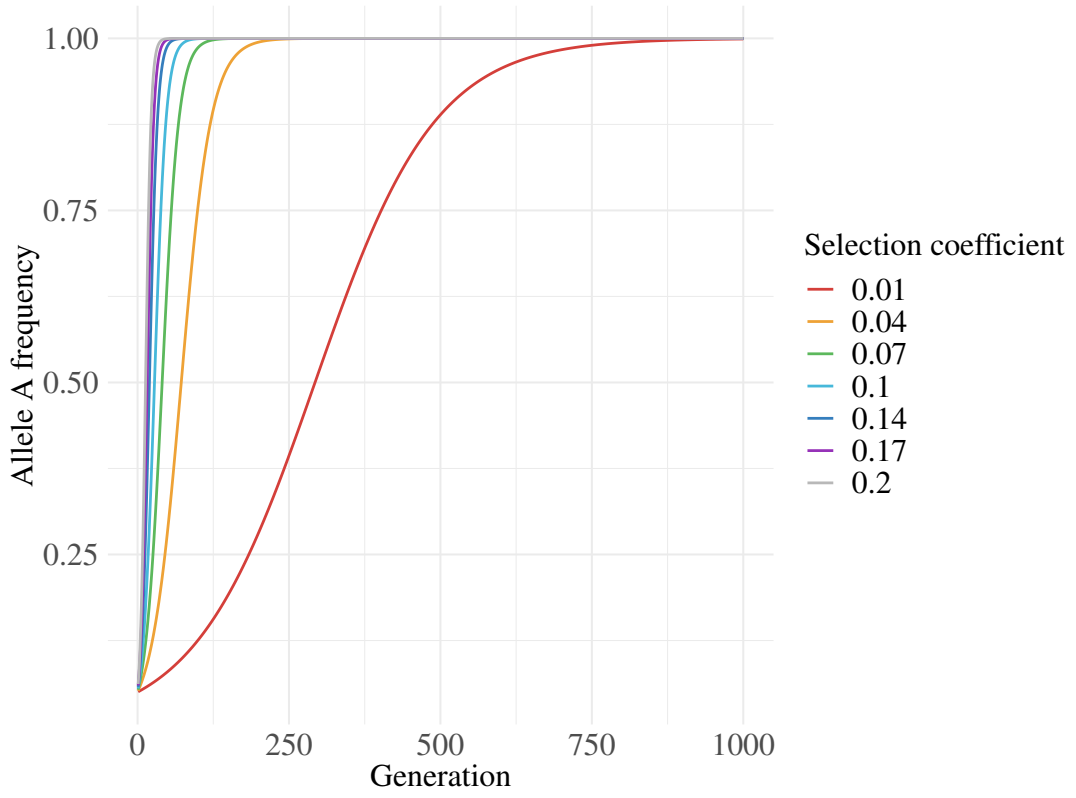

Figure 1.9: Allele frequency changes through time according to s in a haploid population

From figure 1.9, we can also see that the time, for example to go to $f_A = 0.5$ is not a linear function of s. If we solve for t, we get:

$$t = \frac{1}{\log(1-s)} \log \frac{f_A(t=0)/f_A(t) - f_A(t=0)}{f_a(t=0)} \tag{1.67}$$

Therefore the time to significantly change the initial frequency in an haploid population is proportional to $-\frac{1}{\log(1-s)} \approx \frac{1}{s}$ when s is small. In other words, if s is increased by some multiple of s, from s to Ks, then the time needed to reach the desired value is divided by K.

Furthermore we can see from figure 1.9, that for this deterministic model, the speed of allele fixation is maximal at t_{50}, the time at which the allele frequency equal 50%. Hence for initial allele frequency close to the boundaries (0 or 1), the fixation speed tends to 0. Hence, selection effect on a population might be overpowered by other evolutionary forces like, random genetic drift when allele frequencies are close to zero or one.

1.2.4.2 Diploid extension

In the same way, if we consider a biallelic locus in a diploid population, we can compute the change in the genotypic frequency, of, let's say, AA individuals. We have then:

$$f_{AA}(t=1) = \frac{f_{AA}(t=0)w_{AA}}{\overline{w}} \text{ with } \overline{w} = f_{AA}(t=0)w_{AA} + f_{Aa}(t=0)w_{Aa} + f_{aa}(t=0)w_{aa} \tag{1.68}$$

If we suppose that this population reproduces randomly, we can compute the change in allele frequency of A between $t=0$ and $t=1$. We can show that:

$$\Delta f_A = \frac{f_A f_a}{\overline{w}}[f_A(w_{AA} - w_{Aa}) - f_a(w_{aa} - w_{Aa})] = f_A f_a \frac{d\ln\overline{w}}{2df_A} \tag{1.69}$$

From 1.69, two observations can be made:

- The only state of stable polymorphic equilibrium is when the selective value of heterozygotes is greater than that of both homozygotes ($w_{Aa} > w_{aa}$ and $w_{Aa} > w_{AA}$). This special situation is called *overdominance* and is quite rare in nature (Fiévet, 2004).
- Another special case is when the selective value of heterozygotes is lower than that of both homozygotes ($w_{Aa} < w_{aa}$ and $w_{Aa} < w_{AA}$). In that case, called *underdominance*, the frequency of the A allele can increase or decrease, depending on the value of the ratio $\frac{f_A}{f_a}$ as compared to the ratio $\frac{w_{aa}-w_{Aa}}{w_{AA}-w_{Aa}}$.
- In all other cases, the frequency of the most favorable allele will increase until fixation.
- Natural selection leads to the increase of mean population fitness. The speed of change of allelic frequencies are proportional to $\frac{d\ln\overline{w}}{2df_A}$. However, a local maximum can be reached. Then, natural seletion does not guarantee fitness maximization.

Again, in this equation, fitness is viewed as a relative measure of reproductive success and several notations can be used:

AA	Aa	aa
w_{AA}	w_{Aa}	w_{aa}
$1+s$	$1+hs$	1
1	$1-hs$	$1-s$

Table 1.2: Different fitness notations

s is the selection coefficient, h is called the degree of dominance and both s and h are included between 0 and 1. Dominance affects the speed of fixation by changing the relative selective value of heterozygotes, as compared to the two homozygotes.

In the absence of other evolutionary pressures, a favorable allele's fate depends on the selective value of the heterozygotes compared to that of homozygotes. There is a state of polymorphic equilibrium only when the value selective heterozygotes is greater than that of both homozygotes (called overdominance). In all other cases, an allele is lost. So in general, natural selection does not allow to explain the genetic polymorphism observed in populations. However notable exceptions do exists like: frequency dependent selection ($w(f_A)$) and selection varying through time or space ($w(f_A, x, y, z, t)$).

1.2.4.3 Selection from the vantage point of a quantitative trait and the breeder equation

Until now, we adopted the vision of traditional population genetics and focused more on the change of allele frequencies due to evolutionary pressures without much focus on the phenotypic trait. If the previous monolocus models are of considerable importance for discrete traits such as bacterial resistance, quantitative traits that take continuous values are not as well depicted, as they are encoded by many genes along the genome. R.A. Fisher, the father of quantitative genetics, while reconciling the Mendelian school of genetics with the biometricians of his time, proposed a very rich and insightful view of adaptation. We will start by presenting the effect of selection on a trait, without much focus on the underlying genes before making the link with population genetics as Fisher did by considering his famous infinitesimal model.

Fisher's decomposition of phenotypic effects

Indeed, in 1918, Fisher proposed a probabilistic model of decomposition of the phenotypic value. One can consider a quantitative trait Y , as the sum of a genetic effect (G), the average effect of a genotype on its phenotype, and one micro-environmental effects (E). The micro-environmental components of the parents are independent of those of the children in the absence of environmental correlations, and specific to each individual. (Note, however, that E can be transmitted in part for all characters having a "social" component). When two different genotypes respond to environmental variation in different ways, one need also to include $G \times E$ interactions, that we won't elucidate here. Hence one can write:

$$Y = G + E + G \times E \tag{1.70}$$

We will neglect $G \times E$ interactions for this introduction, and only consider:

$$Y = G + E \tag{1.71}$$

Based on Mendelian transmission of genes in a panmictic population, one can further decompose the genetic effects G into additive effects (A) and non additive effects (NA):

$$G = A + NA = A + D + I + W \text{ with } NA = D + I + W \text{ where we have:} \tag{1.72}$$

- **Additive effects (A)** : Each parent transmits to his children half of its additive effects, which corresponds to the fact that a parental gamete, at the end of meiosis, contains half of the genetic information of the parent.
- **Dominance effects (D)** : Dominance effects are interactions between alleles at the same locus. When gametes meet at random, it is impossible to predict the allele received from the mother knowing the allele received from the father. The dominance effect is therefore specific to each individual. Note that this doesn't hold with inbreeding.
- **Epistatic effects (I)** : Epistatic effects are interactions between alleles at different loci. In the presence of genetic linkage, some of the effects of epistasis are transmitted as additive effects.
- **Meiosis effects (W)** : A parent transmitting half of his genes, he transmits in average half of its additive genetic value. However, the transmission results from the random draw of a copy of each gene among the two, the value of the offspring therefore includes a random term.

Notice that all terms above are random variables that can be described by a mean and a variance in a population. $\mathbb{V}(Y)$ is the phenotypic variance in the population. Because in Fisher's model, all random terms A, D, I, W and E are supposed to be independent, we have:

$$\mathbb{V}(Y) = \mathbb{V}(G) + \mathbb{V}(E) \tag{1.73}$$
$$\mathbb{V}(G) = \mathbb{V}(A) + \mathbb{V}(D) + \mathbb{V}(I) + \mathbb{V}(W) \tag{1.74}$$

In a population, each individual has a different phenotype. Phenotypic similarity between relatives can only be approached statistically by the measurement of a *covariance*. However, one can express the phenotype of an offspring subscripted with the letter o, according to the genetic effects of its father (subscripted with f) and its mother (subscripted with m):

$$\left.\begin{array}{l} Y_f = A_f + D_f + I_f + W_f + E_f \\ Y_m = A_m + D_m + I_m + W_m + E_m \end{array}\right\} \Rightarrow Y_o = \frac{1}{2}A_f + \frac{1}{2}A_m + D_o + I_o + W_o + E_o \tag{1.75}$$

Recall that the subscripts chosen for the component of Y_o are the result of many hypotheses:

- Panmixia : Dominance and epistasis effects in the offspring cannot be predicted.
- Linkage equilibrium : Epistasis effect in the offspring result from the random shuffling of allelic combinations at different loci.

- Absence of common environmental effect

Given those hypotheses, the only effects that are transmitted from parent to offspring are the additive effects. The covariance between a father and one of it's offspring writes:

$$\begin{aligned}\mathrm{Cov}(Y_f, Y_o) &= \mathrm{Cov}(A_f + D_f + I_f + W_f + E_f, \frac{1}{2}A_f + \frac{1}{2}A_m + D_o + I_o + W_o + E_o) \\ &= \mathrm{Cov}(A_f, \frac{1}{2}A_f) + \mathrm{Cov}(A_f, \frac{1}{2}A_m)\,\mathrm{Cov}(A_f, D_o) + \mathrm{Cov}(A_f, I_o) + \\ &\mathrm{Cov}(A_f, W_o) + \mathrm{Cov}(A_f, E_o) + \mathrm{Cov}(D_f, \frac{1}{2}A_f) + \mathrm{Cov}(D_f, \frac{1}{2}A_m) + \\ &\mathrm{Cov}(D_f, D_o) + \ \ldots\ + \mathrm{Cov}(E_f, E_o)\end{aligned}$$

Provided the many hypotheses enumerated above, most of the terms are independent leading to a null covariance :

- Panmixia : $\mathrm{Cov}(A_f, A_m) = 0$, $\mathrm{Cov}(A_f, D_o) = 0$, $\mathrm{Cov}(D_f, D_o) = 0$, ...
- Linkage equilibrium : $\mathrm{Cov}(A_f, I_0) = 0$, $\mathrm{Cov}(I_f, I_o) = 0, ...$
- No transmission of environmental effects : $\mathrm{Cov}(A_f, E_0) = 0$, $\mathrm{Cov}(E_f, E_o) = 0$, ...

Thus, remembering that $\mathrm{Cov}(A_f, A_f) = \mathbb{V}(A)$:

$$\mathrm{Cov}(Y_f, Y_o) = \frac{1}{2}\mathbb{V}(A) \tag{1.76}$$

Therefore, the regression line between parent and offspring has a slope $b_{f/o}$ equal to:

$$b_{f/o} = \frac{1}{2}\frac{\mathbb{V}(A)}{\mathbb{V}(Y)} = \frac{1}{2}h^2 \tag{1.77}$$

where h^2 is called the *narrow sense heritability*

With these tools, one can define the widely used concept of heritabilities:

- **Broad sense heritability (H^2)** : The proportion of genetic variance in the phenotypic variance and is computed as:

$$H^2 = \frac{\mathbb{V}(G)}{\mathbb{V}(P)} \tag{1.78}$$

- **Narrow sense heritability (h^2)** : The proportion of additive variance in the phenotypic variance and is computed as:

$$h^2 = \frac{\mathbb{V}(A)}{\mathbb{V}(P)} \tag{1.79}$$

We will be very cautious in their interpretation. Heritability must not be confounded with heredity. Indeed, heritabilities are defined as statistical concepts that are dependent on the population and environment in which there are estimated and do not represent the nature of inheritance. **Heritabilities measure the relative parts of variance components in a population**. A low heritability does not mean the absence of genetic determinism, but the absence of genetic variability for the trait in the population. For example, the character "number of ears" in the human population is genetically determined. The genetic variance is nevertheless very small compared to the phenotypic variance in the population, which is much more influenced by hairdressers ability, or other mishap... Therefore, the heritability of the number of ears in the human population is close to zero.

Indeed, all variance components depend not only on the effect of the genes on the phenotype, but also on allele frequencies. Variance components and heritabilities can be computed for monogenic traits. Figure 1.10 shows the variation of the additive (VarA), dominance (VarD) and total genetic (VarG) variance for trait determined by a single locus with two alleles, A_1 and A_2 as a function of the allele frequency of A_1.

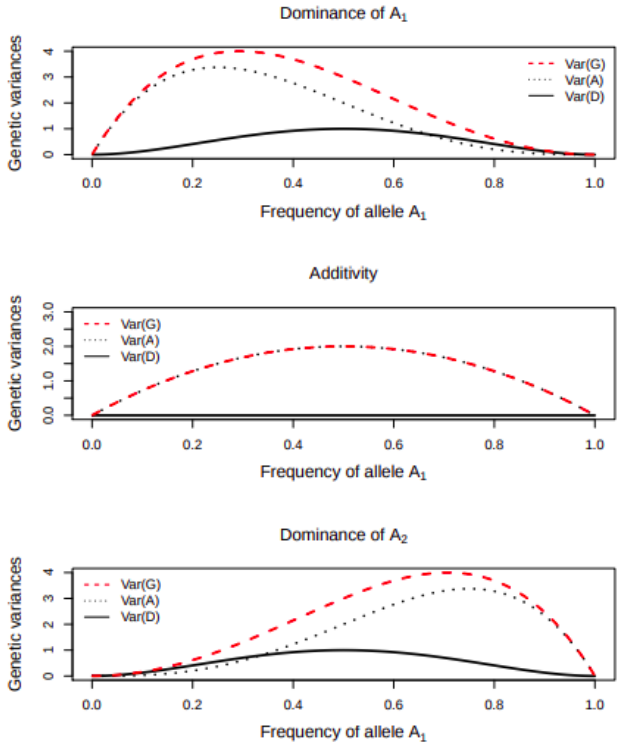

Figure 1.10: Variance components change with allele frequencies in a panmictic population (in Hardy-Weinberg equilibrium) of genotypes with a single biallelic diploid locus. Extracted from Petrizzelli et al. (2019)

Without dominance, the genetic and the additive variance are maximum when the genetic diversity is maximum, *i.e.* when $f_{A_1} = f_{A_2} = 0.5$. When one of the allele is dominant, the additive variance is not null but captures dominance effects: it is maximum when the allele frequency of the dominant allele is not too high. Moreover, because in Fisher's statistical decomposition of genetic effects, dominance terms are computed after adjusting for additive effect, the additive variance captures non-additive genetics effects.

The effect of selection on a quantitative trait As we have seen previously, selection does not modify the traits inheritance, but only the underlying allele frequencies, through the changes in the probability that a parent contributes to the next generation. Let's consider a panmictic population with non overlapping generations, and no effect of sex on the considered phenotype. Let's call X_t the random variable that describes the phenotypic value of an individual of the population at generation t. X_t^f (resp. X_t^m) corresponds to the phenotypic value of the father (resp. mother) of individual X_{t+1}. First, we can use the parent-offspring regression model to describe the phenotypic value of the offspring at generation $t+1$:

$$X_{t+1} = \mu_t + h^2(\frac{X_t^f + X_t^m}{2} - \mu_t) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2) \text{ and } \mu_t = \mathbb{E}(X_t) \tag{1.80}$$

Now, let's consider a fitness function giving the probability that a parent of a given phenotype reproduces: $w(x)$. We first want to compute the mean of the selected parents. This can be achieved by

integrating on all possible parental phenotypic values, weighted by their respective probability of reproduction $w(x)$, and the density $f_t(x)$ of the phenotypic values at time t. We have to normalize it by the integrated product $\bar{w} = \int_{-\infty}^{+\infty} w(x) f_t(x)$, *i.e.* the mean fitness of the population. We have then:

$$\mu_{s,t} = \frac{\int_{-\infty}^{+\infty} x_t w(x) f_t(x) dx}{\int_{-\infty}^{+\infty} w(x) f_t(x) dx} \tag{1.81}$$

We are now interested in the expected mean of the new generation of individuals. We have:

$$\mu_{t+1} = \mu_t + h^2(\mu_{s,t} - \mu_t) \iff \mu_{t+1} - \mu_t = h^2(\mu_{s,t} - \mu_t) \tag{1.82}$$

$\mu_{t+1} - \mu_t$ is the difference between the mean of generation $t+1$ and the mean of the whole parental population, also called **selection response** R:

$$R = \mu_{t+1} - \mu_t \tag{1.83}$$

$\mu_{s,t} - \mu_t$ is the difference between the mean of the selected parents at generation t and the mean of the whole parental population, also called **selection differential S**.

$$S = \mu_{s,t} - \mu_t \tag{1.84}$$

Written this way, we obtain the famous **breeder equation**:

$$R = h^2 S \tag{1.85}$$

In other words, the response to selection is proportionally linked to the difference between selected individuals and the whole population by the narrow sense heritability.

N.B: if we have access to several consecutive generations, several couples (R_t, S_t) can be computed, and we can define the **realized heritability**, as the regression slope of $R = f(S)$.

Selection Types Traditionally, we distinguish several types of selection, characterized by distinct fitness function:
Stabilizing selection: Where selection acts against individual with the most extreme values, the highest and the lowest one, whose fitness is minimal.
Directional selection: Where selection acts in one direction against the most extreme of the phenotypic distribution. One example widely used by breeder is truncation selection, where only individual above (or below) a threshold are allowed to reproduce.
Disruptive selection: Where selection acts against individual with intermediate phenotypic values.

In the particular case of *truncation selection*, $w(x) = 0$ for x lower than the truncation threshold T, and $w(x) = 1$ otherwise, as represented by the red area in (Figure 1.11). Here, S is expressed in the unit of the measured phenotype and difficult to interpret and compare to other selected trait.

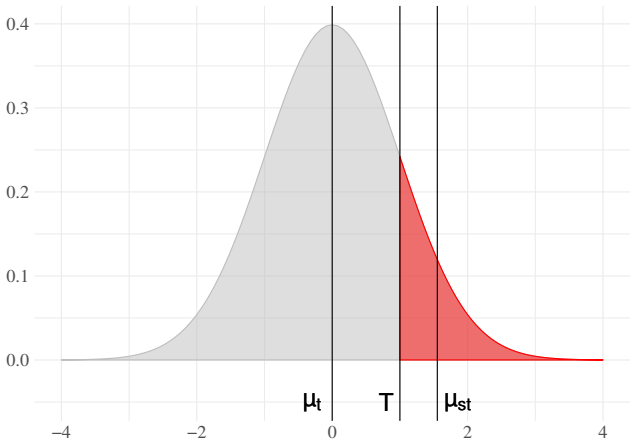

Figure 1.11: **Selection by truncation** Phenotypic values X_t a supposed have a Gaussian distribution in the population. Truncation selection consist in reproducing only the individuals having X_t higher than a given threshold. $\mu_{s,t}$ is the mean of the selected individuals.

Therefore it is often normalized by the standard deviation of the trait: $S = i\sigma_P$ where σ_P is the phenotypic standard deviation, and i is called **selection intensity** and corresponds to the average value of the right-tail of a $Normal(0, 1)$ distribution truncated at $(T - \mu_t)/\sigma_P$.

This lead to an other equivalent way of presenting the **breeder equation**:

$$R = h^2 S = \frac{\sigma_A^2}{\sigma_P^2} i\sigma_P = ih^2\sigma_P \tag{1.86}$$

This writing emphasizes here the fact that the selection response is proportional to the selection intensity, the the narrow sense heritability, and the amount of phenotypic variance.

1.2.4.4 The infinitesimal model

The central assumptions of many quantitative-genetic selection models rely on the hypothesis that the number of loci is assumed to be sufficiently large so that the amount of phenotypic variation attributable to any single locus is small, and hence the amount of selection on any locus is also small. The infinitesimal model have been proposed by Fisher Ronald Aylmer (1930) as the as the limit of Mendelian inheritance, where a phenotypic trait is represented by the total sum of an effectively infinite number of Mendelian loci, each contributing an infinitesimal amount to the total phenotype. It was mathematically formalized by Bulmer and Others (1980), and revisited more recently by Barton (Barton et al., 2017).

Under the infinitesimal model, allele frequencies remains approximately unchanged by selection. Selection acts on the mean phenotypic value of a population by summing infinitesimal allele frequency changes at a large number of loci. To get an intuitive idea of the phenomenon, consider n independent (in linkage equilibrium) biallelic loci with alleles L and l. Let's consider that each locus has the same effects (L having an allelic effects $+a$ and l having an effect equal to zero) and same frequencies (L being in frequency p and l in frequency $1 - p$). Then the expected values for the trait mean and variance are:

$$\mu_A = 2npa \text{ and variance } \mathbb{V}(A) = 2np(1-p)a^2 \tag{1.87}$$

For $\mathbb{V}(A)$ to be bounded when $n \to \infty$, a need to be in order of $\mathcal{O}(\frac{1}{\sqrt{n}})$.

Let's consider a change in allele frequency Δp, then

$$\Delta\mu_A = 2 \cdot n \cdot a \cdot \Delta p \tag{1.88}$$

If a is of order $\mathcal{O}(\frac{1}{\sqrt{n}})$, $\Delta\mu_A$ will be of order $\mathcal{O}(\sqrt{n} \cdot \Delta p)$. If n in large enough, this may correspond to an observable change in the trait mean value.

Under the infinitesimal model, the mean value of a trait can change under selection without noticeable allele frequency changes.

It easily follows that after one generation of selection:

$$\Delta\,\mathbb{V}(A) = 2n\Delta p(1-2p-\Delta p) \tag{1.89}$$
$$\approx a(1-2p)\Delta\mu \tag{1.90}$$

Since a is of the order of magnitude of $\mathcal{O}(\frac{1}{\sqrt{n}})$, $\Delta\,\mathbb{V}(A)$ is of the order of magnitude of $\mathcal{O}(\Delta p)$ which is close to zero.

Under the infinitesimal model and linkage equilibrium, selection doesn't change the additive genetic variance.

1.2.4.5 Selection gradient

Let's return to the selective value $w_t(x)$ associated to phenotype x. Assume that this function is continuous and integrable for all values of x. The mean fitness of the population is:

$$\overline{w_t} = \int w_t(x)f(x)dx \tag{1.91}$$

As we have seen, selection will lead to a mean fitness change, due to a mean trait change. Let's presume that $X_t \sim \mathcal{N}(\mu_t, \mathbb{V}(P))$. We want to quantify this mean fitness change, due to a mean trait change. In other words, we are interested in assessing $\frac{d\overline{w_t}}{d\mu_t}$, but to do so we need to use the fact that the fitness of an individual $w_t(x)$ is independent of the phenotypic mean μ_t of the population and we need to use the following property:

$$\frac{df(x_t)}{d\mu_t} \approx f(x_t)\frac{x_t-\mu_t}{\mathbb{V}(P)} \tag{1.92}$$

One find then:

$$\frac{d\overline{w_t}}{d\mu_t} = \frac{1}{\mathbb{V}(P)}\left[\int x_t w_t(x_t)f(x_t)dx_t - \mu_t\int w_t(x)f(x_t)dx_t\right] \tag{1.93}$$
$$= \frac{1}{\mathbb{V}(P)}\left[\overline{w_t}\mu_{s,t} - \mu_t\overline{w_t}\right] \tag{1.94}$$

$\Longleftrightarrow$

$$\mu_{s,t} - \mu_t = \mathbb{V}(P)\frac{d\ln\overline{w_t}}{d\mu_t} \tag{1.95}$$

One can then rewrite the breeder equation as:

$$R = \mu_{t+1} - \mu_t = \mathbb{V}(A)\frac{d\ln\overline{w_t}}{d\mu_t} \tag{1.96}$$

where $\frac{d\ln\overline{w_t}}{d\mu_t}$ is called the **selection gradient**. It is measured as the slope of the function linking the population mean selective value to the population phenotypic mean and reflects the strength of selection on the trait. This formulation, and its extension to the multivariate case, was initially proposed by Lande (1976, 1979). It highlights three observations:

- The response to selection on a trait depends on both the additive variance and the strength of selection.
- Selection will move the population phenotypic mean towards a local optimum of the fitness function.
- The applications of quantitative genetics are not reduced to breeding programs and also cover the study of evolution in natural populations.

1.2.4.6 Fitness landscape definition

A famous visualization of the relationship between fitness and genotypes has been promoted by Wright (1932): the concept of fitness landscape. Indeed, if we consider fitness as a function defined on a multidimensional genotypic space, we can represent it on a diagram with one axis representing the fitness value and one or two axes as projection of the genotypic space, as shown figure 1.12. Fitness landscapes are often depicted as ranges of mountains, where points from which all paths are downhill, are called local adaptive peaks, and regions from which many paths lead uphills are called valleys. A fitness landscape with many peaks and valleys is called rugged. In this landscape, the distance between two points can be seen as an evolutionary distance.

Hence, adaptation can be seen as the gradual walk of a population towards an adaptive peak.

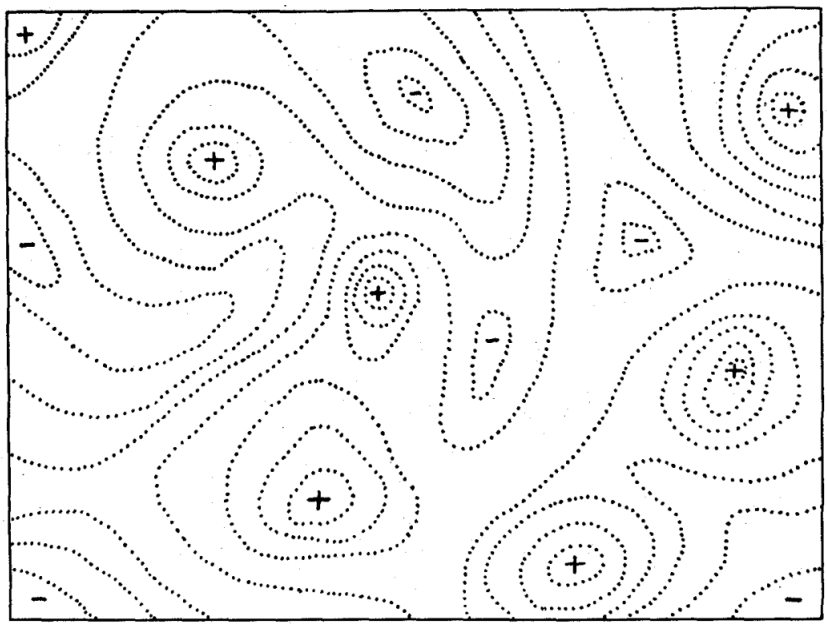

FIGURE 2.—Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.

Figure 1.12: **S. Wright's visualization of a fitness landscape**, (Wright, 1932). The two dimensions represent two phenotypic traits and the contour lines the fitness.

Three main fitness landscape types have been used:

- Genotypes fitness landscapes: Fitness is considered as a function defined on a multidimensional genotypic space. This space is not continuous. Instead each points is linked to another through a mutational network.
- Allele frequencies fitness landscapes: Fitness is considered as a function of allele frequencies.
- Phenotypes fitness landscapes: Fitness is considered as a function defined on a space of phenotypic traits.

Let's consider a phenotypes fitness landscape. If all the traits that determine fitness have Gaussian distributions, we can use the properties of the multivariate normal distribution to extend Lande's equation to the multitrait case. For example for two taits A and B, we will have:

$$\begin{pmatrix} \Delta\mu^A \\ \Delta\mu^B \end{pmatrix} = \begin{pmatrix} \mathbb{V}(A^A) & \mathrm{Cov}(A^A, A^B) \\ \mathrm{Cov}(A^A, A^B) & \mathbb{V}(A^B) \end{pmatrix} \begin{pmatrix} \beta^A = \frac{\partial \ln \overline{w_t}}{\partial \mu^A} \\ \beta^B = \frac{\partial \ln \overline{w_t}}{\partial \mu^B} \end{pmatrix} \tag{1.97}$$

$$\Delta \overrightarrow{\mu} = G \overrightarrow{\beta} \tag{1.98}$$

The β coefficients are called **selection gradient** and determine the adaptive landscape. G is the variance-covariance matrix of additive effects. It is called **G** matrix. By determining the strongest axes of genetic variation, the matrix G will determine the direction of phenotypic evolution in a given adaptive landscape.

1.2.5 Other evolutionary forces

In this thesis, I chose to detail the evolutionary forces that will be of particular interest for my work. Two other evolutionary forces are of importance in evolutionary biology: migration and mating system. I briefly review here some fundamentals and will develop self-fertilisation (after introducing the concept of effective population size which is central when considering interacting evolutionary forces).

1.2.5.1 Migration

Migration brings new alleles into a population. From a population genetics point of view, its effect are similar to those of mutations: a migrant arrives with its haplotypes and it is a source of linkage disequilibrium. In diploids, effects of migration on allele frequencies depend on whether the migrants arrived as diploids (the individuals are moving) or haploids (gametes or gametophytes are moving). Also, migration rates can be much higher than mutation rates.

At one biallelic locus, the admixture of two panmictic populations results in a deficit in the frequency of heterozygotes, as compared to Hardy-Weinberg proportions. This effect, also known as Wahlund effect (Wahlund, 1928). , has led to the concept of F-statistics that measures the differentiation between and within populations (Wright, 1949).

More generally, migration is associated with the concept of metapopulations: the individuals of the same species form a network of populations that are interconnected by migration (Levins, 1969; Couvet et al., 1985).

1.2.5.2 Mating systems

In sexual eukaryotes, the mating system describes the encounter of gametes or gametophytes prime to fecundation. Mating systems typically change the relation between allelic and genotypic frequencies. The reference mating system is panmixia or random mating, that results, within a single generation in the Hardy-Weinberg equilibrium. The mating system can be determined by single loci or chromosomes like in the case of auto-incompatibility or sex determination, or by phenotypic traits like in the case of homogamy (preference mating between similar individuals) or heterogamy (preference mating between dissimilar individuals). In all cases, it results either in an excess, or in a deficit of heterozygotes.

1.3 Interaction between evolutionary forces

Evolutionary forces rarely act alone. Rather, genetic and phenotypic diversity that can be observed in a population at a given generation are the result of the combination of evolutionary forces during many generations. Here, I detail theoretical advances that allowed to better understand interactions between evolutionary forces.

1.3.1 Effective population size: Ne

Wright-Fisher model is very convenient for its mathematical properties but is based on an idealized view of populations. However, most populations deviate from these hypotheses, especially when they encounter several interacting evolutionary forces. One way to circumvent this issue would be to design more realistic models, like Moran models (Moran, 1958) allowing for overlapping generations and used to model birth-death processes. The results obtained with Moran models, however, often converge to the ones obtained with Wright-Fisher model, as N increases.

In 1931, S.W. Wright suggested an other approach: the concept of effective population size, which became essential in population genetics. In simple terms, **the effective population size is equivalent to the number of individuals that an ideal population (like a Wright-Fisher population or a Moran population) would require to get the same behavior as the focal population**. The different factors affecting N_e have been listed in Charlesworth (2009):

- Variation in offspring number per individual that does not follow a binomial distribution.
- Inbreeding that tends to decrease N_e.
- Changes in population size: for example, population bottlenecks greatly reduce N_e.
- Genetic structure and selection: the long-term maintenance of two or more alleles by balancing selection results in an increase in N_e at sites that are closely linked to the target of selection. In contrast, directional selection causes a reduction in Ne at linked sites. This process correspond to the Hill–Robertson effect.
- Inheritance mode: for example, autosomal, X-linked, Y-linked or organelle.
- Separate sexes and differences in male and female census size (decrease N_e resulting from this mechanism is encountered in bovine breeding programs).
- Age- and stage-structured populations.
- Spatial structure and metapopulations.

Yet, the many ways to depart from an *idealized* population do not affect the different properties in the same way. This led to many different interpretations, and different methods for the estimation of N_e (Kimura and Crow, 1963; Crow and Kimura, 1971).

Inbreeding effective size (N_e^f): Let's consider a single locus in an ideal monoecious diploid population, reproducing trough panmixia, with possible selfing, and no selection. A progeny is produced by randomly drawing with replacement a pair of alleles in an infinite pool of gametes. Let's consider the inbreeding coefficient, the probability that two alleles at any locus in an individual are identical by descent, f_t at generation t.

$$f_t = \frac{1}{2N_{t-1}} + \left(1 - \frac{1}{2N_{t-1}}\right) f_{t-1} \tag{1.99}$$

It corresponds to the probability that the two alleles of the progeny come from the same parental allele ($\frac{1}{2N_{t-1}}$) plus the probability that the two drawn alleles don't $1 - \frac{1}{2N_{t-1}}$ but were nevertheless identical by descent in the previous generation, f_{t-1}.

Now let's consider a non ideal population, and call P_t the probability that two uniting gametes comes from the same parent. Then we have:

$$f_t = \frac{P_t}{2} + \left(1 - \frac{P_t}{2}\right) f_{t-1} \tag{1.100}$$

Indeed, the probability that the two alleles of the progeny comes from the same parental allele is equal to the probability that they come from the same parent P_t times the probability that the probability that they are the same 1/2.
Then, the two equation are equivalent if:

$$P_t = \frac{1}{N_t} \tag{1.101}$$

Therefore, one can define the inbreeding effective population size as the inverse of the probability that two randomly chosen gametes come from the same parent.

$$N_e^f = \frac{1}{P_t} \tag{1.102}$$

Hence this effective population size definition relies on the number of individuals in the parental population.

Now, let's try to compute P_t more generally. Let's call $k_{t,i}$ the random variable giving the number of gametes of individual i at time t, with mean $\overline{k}$ and variance V_k. The number of ways in which two gametes can be chosen out of the total number of gametes $N_{t-1}\overline{k}$, is $\binom{N_{t-1}\overline{k}}{2}$ of which $\sum_i \binom{k_{t-1,i}}{2}$ is the number of cases in which two gametes come from the same parent. Hence,

$$P_t = \frac{\sum_i \binom{k_{t-1,i}}{2}}{\binom{N_{t-1}\overline{k}}{2}} = \frac{\sum_i k_{t-1,i}(k_{t-1,i}-1)}{N_{t-1}\overline{k}(N_{t-1}\overline{k}-1)} \tag{1.103}$$

After noting that,

$$\overline{k} = \frac{\sum_i k_{t-1,i}}{N_{t-1}} \text{ and } V_k = \frac{\sum_i {k_{t-1,i}}^2}{N_{t-1}} - \overline{k}^2 \text{ and } N_{t-1}\overline{k} = 2N_t \tag{1.104}$$

it follows,

$$P_t = \frac{V_k + \overline{k}(\overline{k}-1)}{\overline{k}(2N_t - 1)} \tag{1.105}$$

Finally, we have:

$$N_e^f = \frac{2N_t - 1}{\overline{k} - 1 + V_k/\overline{k}} \tag{1.106}$$

If the population size remains constant, $\overline{k} = 2$, and N_e^f depends on the census population size and on the variance of offspring numbers V_k

$$N_e^f = \frac{4N - 2}{V_k + 2} \tag{1.107}$$

The inbreeding effective size is the size of an ideal population that would have the same inbreeding coefficient as the observed one.

Variance effective size (N_e^v): In our previous diploid Wright-Fisher model, we have seen that:

$$\mathbb{V}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N} \tag{1.108}$$

If we estimate $\mathbb{V}(p_{t+1}|p_t)$ in a sampled population, $\widehat{var(p_{t+1}|p_t)}$ is typically larger than the former variance, but if we replace the former by the later in the above formula and solve for N, we can derive:

$$N_e^v = \frac{p_t(1-p_t)}{\widehat{var(p_t)}} \tag{1.109}$$

Using the same kind of reasoning as for the inbreeding effective population size, Crow and Kimura (1971) derived the following expression:

$$N_e^v = \frac{(2N_{t-1}-1)\overline{k}}{2(1+V_k/\overline{k})} \tag{1.110}$$

Here the variance effective size is related to the progenies' population size.

If the population size remains constant, $\overline{k}=2$, hence:

$$N_e^v = \frac{4N-2}{V_k+2} \tag{1.111}$$

The variance effective size is the size of an ideal population that would have the same variance of allele frequencies from one generation to the next as the observed one.

We have seen, that if the population size remains constant, N_e^f and N_e^v are equivalent. However, they won't be if the population size changes. Consider two extreme cases: firstly if one individual gives rise to many offspring (*i.e.* $N_{t-1}=1$ and $\overline{k}\to\infty$), $N_e^f=1$ while $N_e^v\to\infty$; secondly, if each parent gives rise to only one child, (*i.e.* $\overline{k}=1$ and $V_k=0$), $N_e^f\to\infty$ but N_e^v stays finite.

Coalescent effective size (N_e^c): We can use the derived parameters of the coalescent theory to estimate N_e^c. If we place ourselves under the infinite site model, we can define the nucleotide diversity π as the probability that two nucleotides differ in two randomly drawn individuals, which corresponds to the probability of mutation before coalescence, which is equal to

$$\pi = \frac{\theta}{1+\theta} \approx \theta = 4N_e^c\mu \tag{1.112}$$

Therefore, knowing the mutation rate and the nucleotidic diversity, one can estimate directly the effective population size from the nucleotide site diversity.

$$N_e^c = \frac{\pi}{4\mu} \tag{1.113}$$

The coalescent effective size is the size of an ideal population that would have the same nucleotidic diversity as the observed one, given the mutation rate.

If we recall, that the time going back to their most recent common ancestor for an haploid Write-Fisher population, is equal to $2N(1-1/k)$ in a sample of k individuals, then we can also estimate N_e^c from the estimation of the T_{MRCA}.

Variation in population size through generations: Setting N constant through time, we have seen that $\mathbb{E}(H_t)=H_0(1-\frac{1}{N})^t$, but if we allow for fluctuations for N_t we can write:

$$\mathbb{E}(H_t)/H_0 = \prod_{k=0}^{t-1}(1-\frac{1}{N_k}) = (1-\frac{1}{N_e})^t \tag{1.114}$$

This equation can be approximated if the N_k are large enough and t small:

$$1 - \frac{t}{N_e^d} = t - \sum_{k=0}^{t-1} \frac{1}{N_k} \tag{1.115}$$

or

$$\frac{1}{N_e^d} = \frac{1}{t} \sum_{k=0}^{t-1} \frac{1}{N_k} \tag{1.116}$$

Thus, if population size changes through time, the effective population size can be approached by the harmonic mean of all population sizes. The demographic effective size is the size of an ideal population that would have the same decrease of Nei's diversity index as the observed one.

See here that, N_e^d is highly impacted by the smallest values of N_k, i.e. by demographical bottleneck.

1.3.2 Self-fertilization

Autogamy, or selfing, refers to the fusion of two gametes that come from one individual. Population geneticists defined the widely used selfing coefficient, σ, the probability of self-fertilization in a population that can vary widely between species (Goodwillie et al., 2005). Following Julien (2019) thesis introduction (publicly available in the coming weeks), autogamy, or selfing, is characterised by the decrease in heterozygosity throughout the genome because of the non independent sampling of gametes during fecondation (Caballero and Hill, 1992). In the extreme case in which a population reproduce exclusively through selfing ($\sigma = 1$), the proportion of heterozygous individual at one locus H_{Obs} decreases as:

$$H_{Obs}(t) = H_{Obs}(0) \times \left(\frac{1}{2}\right)^t \tag{1.117}$$

(Crow and Kimura, 1971). Or in other words, at one locus, initial heterozygozity is expected to be divided by two at each generation of selfing.

This effect leads to a reduction in the effective population size (Fig. 1.13). But, one first can try to compare this deficit in heterozygozity to the expected heterozgosity in a panmictic population. This led to the definition of the F-statistic (Wright, 1949):

$$F_{IS} = 1 - \frac{H_e}{H_{Obs}} \tag{1.118}$$

, where H_e is the expected heterozygosity in panmixia ($2p(1-p)$, with p the allele frequency)

Armed with this definition, Pollak (1988) linked this inbreeding coefficient to the inbreeding population size:

$$N_e^f = \frac{N}{1 + F_{IS}} \tag{1.119}$$

Hence, in complete selfing ($\sigma = 1$), *i.e.* $F_{IS} = 1$, the effective population size is essentially divided by two, compared to an allogamous population. Therefore, we expect an increase of the impact of drift compared to panmixia (Fig. 1.13).

Note, that this increase in homozygosity reduces the impact of recombination, *i.e* a decrease in effective recombination rate r_e (Golding and Strobeck, 1980; Nordborg, 2000). So that :

$$r_e = (1 - F_{IS}) \times r \tag{1.120}$$

Indeed, if all loci are homozygous, recombination cannot change haplotypes. In other words, selfing creates linkage disequilibrium that cannot be removed. This effect can be seen very clearly when comparing the extent of linkage disequilibrium in *Arabidopsis thaliana*, a highly autogamous plant in which LD was detected over 100kb (Nordborg et al., 2005), while Tenaillon et al. (2001) showed LD limited to few

hundreds of bases in allogamous maize. This less efficient recombination leads to an increase in genetic hitchhiking (Maynard Smith and Haigh, 1974) when selecting for a favorable allele (selective sweep), and to background selection when counterselecting for a deleterious mutation. These two effects of selection further deplete neutral genetic diversity (Barton, 2000), especially in autogamous populations (Fig. 1.13).

However, unlike the case of allogamy, a new population can emerge from a single self-pollinated individual, playing a central role in colonisation events of new habitats (Baker, 1967). However, such foundation events have drastic effects on adaptive capacity of autogamous populations , due to extreme genetic diversity decrease, high drift (Whitlock and Barton, 1997).

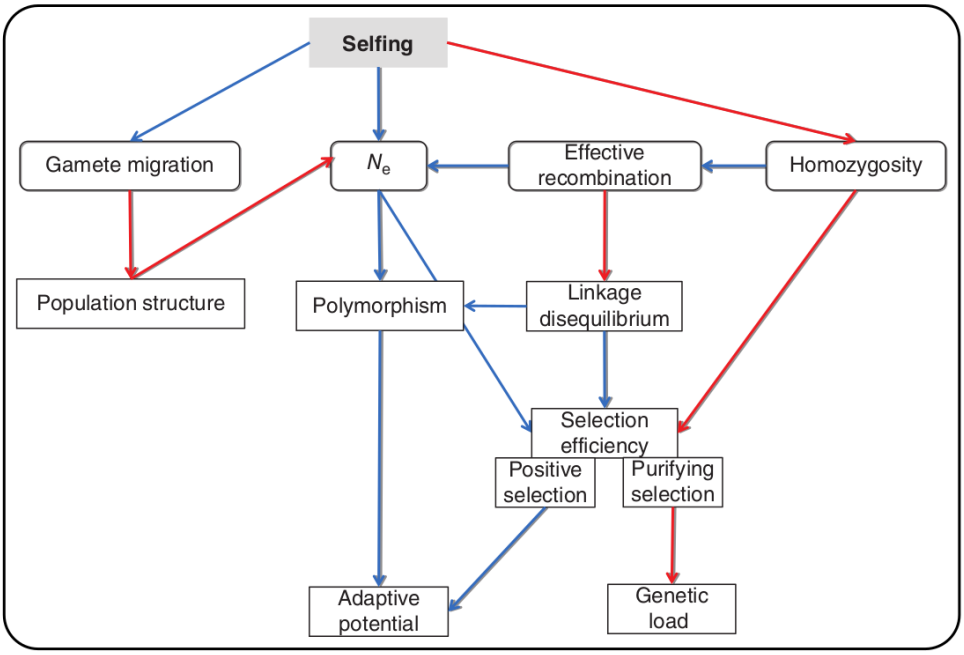

Figure 2 Factors affecting genetic diversity and potential for adaptation in self-fertilising species. Arrows represent cause–effect relations in a decreasing (blue) or increasing (red) sense. N_e is the species effective population size and is globally reduced by selfing.

Figure 1.13: **Schematic summary of selfing on adaptive capacity of autogamous population** Extracted from Burgarella and Glémin (2017)

1.3.3 Drift and Mutation : molecular evolution

Interactions between drift and selection have been widely studied to predict the patterns of molecular evolution. As shown previously, random genetic drift leads to coalescence of pedigrees, and all individuals of a sample are expected to descend from a single ancestor, the $MRCA$. All mutations that occurred in the ancestral path of the $MRCA$ are fixed within populations, but the mutations that occurred on the different lineages descending from the $MRCA$ constitute the genetic diversity of the population.

Let us consider a diploid population constituted of N_e individuals. Considering a mutation rate μ, we expect μr mutations in r generations and if we measure time in unit of $2N_e$ generations, where $t = \frac{r}{2N_e}$, we would expect $2N_e\mu t$ mutations on a lineage of length t. The number of mutation separating two lineages is then twice that, *i.e.* $4N_e\mu t$. We can now define the widely used parameter $\theta = 4Ne\mu$ that measures the nucleotidic diversity. Different methods have been proposed to estimate θ, that rely on two modeling approaches : the infinite alleles model, or the infinite sites model.

1.3.3.1 Infinite alleles model

This model was suggested by Crow and Kimura (1971), and the following ideas are extracted from their papers. This model suggests that the number of possible alleles is so great that each mutation

always corresponds to a new allele. To illustrate this assumption, we can compute the number of possible sequences for a gene of length 4000bp (mean size of a maize gene): as $4^{4000} = 10^{4000\frac{\log 4}{\log 10}} = 10^{2408}$, among which $3 \times 4000 = 12000$ can be reach by a one base change. Therefore, the probability that a mutation occurs at the same loci is $\frac{1}{12000}$, in this example. The number of possible alleles is essentially infinite.

Inbreeding coefficient In a panmictic population of effective size N_e, Wright (1931) and Malécot (1948) showed that F_t, the inbreeding coefficient in generation t, or in other words the probability of identity by descent of two alleles without mutation can be written as:

$$F_t = \frac{1}{2N_e} + \frac{2N_e - 1}{2N_e} F_{t-1} \tag{1.121}$$

If we include mutations, *i.e.* with $(1-\mu)^2$ being the probability that neither allele is mutated at the previous generation, it becomes:

$$F_t = [\frac{1}{2N_e} + \frac{2N_e - 1}{2N_e} F_{t-1}](1-\mu)^2 \tag{1.122}$$

As we have seen, drift leads to the loss of allelic diversity and mutation is the only force producing new alleles. Therefore, an equilibrium, the "mutation-drift equilibrium", might be possible. If we set, $F_t = F_{t-1} = F$, neglect μ^2 terms, and solve for F, we obtain:

$$F \approx \frac{1 - 2\mu}{4N_e\mu - 2\mu + 1} \approx \frac{1}{4N_e\mu + 1} \approx \frac{1}{\theta + 1} \tag{1.123}$$

F can here be seen as the probability that a random individual will be homozygous, or the probability that two randomly chosen alleles will be identical by descent. If $4N_e << \frac{1}{\mu}$, F reaches 1 and all alleles are identical by descent. If $4N_e >> \frac{1}{\mu}$ more alleles are expected to be maintained in the population, and as Ne increases the number of heterozygotes gets larger.

Number of alleles One might be interested in knowing the number of different alleles that are present in a population. If we look backwards from k lineages, and consider the time between k and $k-1$ lineages, there are two possibilities : either a mutation occurs, that will increase the number of alleles by one, or a coalescence event occurs. A coalescence event occurs with probability $\frac{k(k-1)}{2}\frac{1}{2N}$ as seen before. A mutation event occurs with probability $k\mu$. If we measure time in units of $2N$ generations, coalescence occurs at rate $\frac{k(k-1)}{2}$ and mutation at rate $\frac{k\theta}{2}$. Therefore the probability that a mutation occurs first is:

$$\frac{\frac{k\theta}{2}}{\frac{k\theta}{2} + \frac{k(k-1)}{2}} = \frac{\theta}{\theta + k - 1} \tag{1.124}$$

and the probability that a coalescence event occurs first is:

$$\frac{\frac{k(k-1)}{2}}{\frac{k\theta}{2} + \frac{k(k-1)}{2}} = \frac{k-1}{\theta + k - 1} \tag{1.125}$$

Using this reasoning, Watterson (1975) derived an approximate value for the expectation of the number of alleles K_n found in a sample of size n:

$$\mathbb{E}(K_n) \sim \theta \log(n) \tag{1.126}$$

This gives us an asymptotic estimator of θ: $\frac{K_n}{\log n}$, called the Watterson estimator. Unfortunately, the variance of K_n is high ($\mathbb{V}(K_n) \sim \theta \log(n)$), and the convergence to the approximates value is very slow (in times proportional to $\frac{1}{\log n}$).

Ewens (1972) refined this expression by giving the explicit distribution of allelic occurence in a sample of size n, also know as Ewen's sampling formula:

$$\Pr_{\theta,n}(a_1, ..., a_n) = \frac{n!}{\theta(\theta+1)...(\theta+n-1)} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!} \tag{1.127}$$

where $\Pr_{\theta,n}(a_1, ..., a_n)$ is the probability of observing a_1 allele present one time, ..., a_i alleles present i times, to a_n allele present n times. Using this distribution, it is possible to derive an estimate of θ from the observed number of alleles in a sample:

$$\mathbb{E}(K_n) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + ... + \frac{\theta}{\theta+n} \tag{1.128}$$

1.3.3.2 Infinite sites model

The infinite alleles model was developed and greatly studied at a time were genetic information was gathered through indirect means and only few loci were therefore considered. With the advance of genomics, and the availability of whole genome sequences, the other model also developed by Kimura (1969), the infinite sites model, became more natural to study. In this approach, the main hypothesis is that a mutation always occurs at distinct nucleotide/site.

In this framework, we can first try to consider the number of segregating variant, S_n in a sample of size n. We showed previously, that if N is large and time is measured in unit of $2N$ generations, the amount of time in the coalescent during which there are j lineages, $t_j \sim Exp(\frac{j(j-1)}{2})$. If we look at the whole tree, and compute its total length, $T_{tot} = \sum_{j=2}^{n} jt_j$, and take its expected value:

$$\mathbb{E}(T_{tot}) = \sum_{j=2}^{n} j\,\mathbb{E}(t_j) = \sum_{j=2}^{n} j\frac{2}{j(j-1)} = 2\sum_{j=2}^{n}\frac{1}{j-1} = 2\sum_{j=1}^{n-1}\frac{1}{j} \tag{1.129}$$

As mutation occurs at a rate $2N_e\mu$,

$$\mathbb{E}(S_n) = 2N\mu\,\mathbb{E}(T_{tot}) = \theta\sum_{j=1}^{n-1}\frac{1}{j} \tag{1.130}$$

Knowing that the harmonic series converge to $\log(n)$ when n goes to infinity, the infinite sites model and the infinite allele model become equivalent for large n. One can also prove that:

$$\mathbb{V}(S_n) = \theta\sum_{j=1}^{n-1}\frac{1}{j} + \theta^2\sum_{j=1}^{n-1}\frac{1}{j^2} \tag{1.131}$$

To compute the distribution of S_n is a bit more difficult. But we can try to look at a sample of size 2. In the previous paragraph, we have shown that the probability that a mutation occurs before a coalescence event is (replacing by k=2): $\frac{\theta}{\theta+1}$ and that the probability that a coalescence event occurs first is $\frac{1}{\theta+1}$. Armed with these, we can compute the probability that $S_2 = 1$, *i.e.* the probability that a mutation occurred first and then a coalescence event:

$$\Pr(S_2 = 1) = \frac{\theta}{\theta+1}\frac{1}{\theta+1} \tag{1.132}$$

$\Pr(S_2 = 2)$ corresponds to the probability that two mutations occurred before a coalescence event, in this case:

$$\Pr(S_2 = 2) = \left(\frac{\theta}{\theta+1}\right)^2\frac{1}{\theta+1} \tag{1.133}$$

Iterating we obtain:

$$\Pr(S_2 = i) = \left(\frac{\theta}{\theta+1}\right)^i\frac{1}{\theta+1} \tag{1.134}$$

Tavaré (1984) refined this expression by giving the explicit distribution of the number of segregating sites in a sample of size n, S_n:

$$\Pr(S_n = k) = \frac{n-1}{\theta}\sum_{i=1}^{n-1}(-1)^{i-1}\binom{n-2}{i-1}\left(\frac{\theta}{\theta+i}\right)^{k+1} \tag{1.135}$$

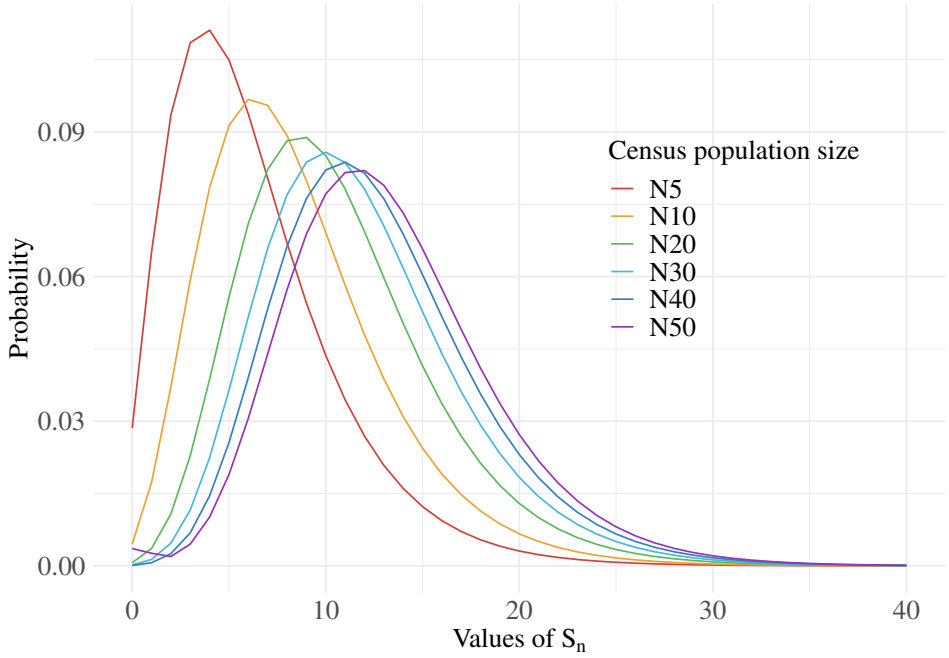

Figure 1.14: Probability distribution of S_n. The distribution tends to a normal distribution as n increases.

1.3.3.3 Unfolded and Folded site frequency spectrum (SFS)

The previous distribution constitutes a great insight on how much sanding variation we would expect to see under mutation and drift alone. However, it gives little information on how it is distributed among individuals of the same population. An alternative description, widely used among population geneticists, is the site frequency spectrum. But first we need to define an *ancestral allele* as the allele of the most recent common ancestor of the sample, and a derived allele. For single nucleotide polymorphisms (SNPs) one can attribute 0 to the ancestral state, and 1 to the derived alleles. For example, for n=5 and 3 polymorphic sites, one can write:

Sequences	Site 1	Site 2	Site 3
Ind 1 (ancestral)	0	0	0
Ind 2	0	1	1
Ind 3	0	1	1
Ind 4	1	0	1
Ind 5	1	0	1

Then we compute the allele frequencies of the derived alleles (here 2/5, 2/5 and 4/5). And after that, one can compute the proportion of derived alleles with frequencies equal to 1/n, 2/n, ... n-1/n. In the example, 0/3 derived alleles have a frequency of 1/5, 2/3 derived alleles have a frequency of 2/5, 0/3 derived alleles have a frequency of 3/5, and 1/3 derived alleles have a frequency of 4/5. Hence the definition of the *unfolded site frequency spectrum*, **f**:

$$\mathbf{f} = (f_1, f_2, ..., f_{n-1}) \tag{1.136}$$

where f_1 is the proportion of sites whose derived alleles have the lowest frequency possible 1/n (in the example $f_1 = 0$) and are called singletons, f_i is the proportion of sites whose derived alleles are in frequency i/n, etc. ($\mathbf{f} = (0, 2/3, 0, 1/3)$)

What is the expected distribution of the folded SFS under drift and mutation alone ? First, let's consider f_1 the proportion of singletons, (in frequency $1/n$). If a mutation occurs on the external branches of Kingman coalescent tree, it will result in a singleton. However, if it happens on the internal branches of the tree, the mutation will be shared by at least two individuals. We can show that the expected length

of the external branches of the tree is equal to 2, (with time in unit of 2N generation). As we would expect $\frac{\theta}{2} \times t$, mutations on a branch of length t, the expected number of singleton is only θ. Therefore:

$$\mathbb{E}(f_1) = \frac{\theta}{\theta \sum_{j=1}^{n-1} \frac{1}{j}} = \frac{1}{\sum_{j=1}^{n-1} \frac{1}{j}} \tag{1.137}$$

Griffiths and Tavaré (1998) used similar considerations to find the expected unfolded SFS, under a infinite site coalescent model:

$$\mathbb{E}(f_i) = \frac{\frac{1}{i}}{\sum_{j=1}^{n-1} \frac{1}{j}} \text{ with } i \in [\![1, n-1]\!] \tag{1.138}$$

However, we see here that the definition of the unfolded site frequency spectrum is dependent on the knowledge of the ancestral state, but this is rarely possible. One might want to bypass this problem, by choosing an outgroup, an individual genetically distant from the sample (*e.g.* an individual of a closely related species), with the hypothesis that the outgroup is closer to the MRCA. However, one can also redefine the site frequency spectrum by *folding* it. This operation consists in defining $\mathbf{f}*$, by adding together frequencies of the derived alleles and the ancestral alleles, by setting:

$$f_i^* = \begin{cases} f_i + f_{n-i} & \text{if } i < \frac{n}{2} \\ f_i & \text{if } i = \frac{n}{2} \end{cases} \tag{1.139}$$

In our example, $\mathbf{f}^* = (f_1^* = 0 + 1/3, f_2^* = 2/3 + 0)$
Then the expected $\mathbf{f}*$, can be computed as:

$$\mathbb{E}(f_i^*) = \begin{cases} \frac{1/i + 1/(n-i)}{\sum_{j=1}^{n-1} 1/j} & \text{if } i < \frac{n}{2} \\ \frac{1/i}{\sum_{j=1}^{n-1} 1/j} & \text{if } i = \frac{n}{2} \end{cases} \tag{1.140}$$

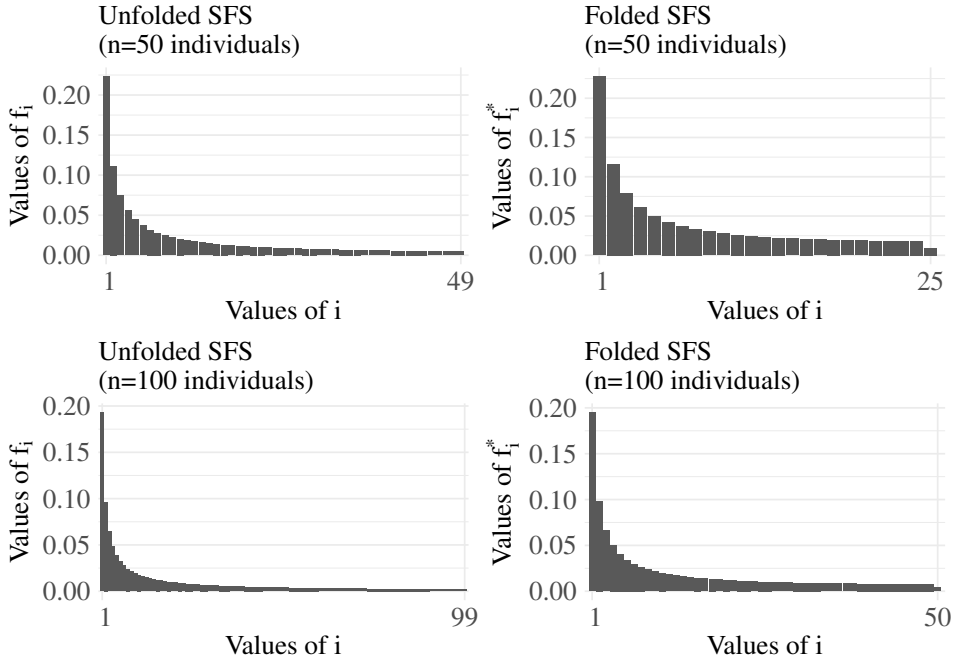

Figure 1.15: Expected neutral folded and unfolded SFS under Kingman coalescent and infinite sites model

The SFS is now widely used in molecular population genetics to compare observations to the expectations under different demographic (*e.g.* demographical expansion in humans (Marth et al., 2004; Gutenkunst et al., 2009)) or mutational scenarios (*e.g.* inference of the distribution of mutational effects in humans (Boyko et al., 2008)).

1.3.4 Interaction between selection and mutation

Whenever a new mutation has a selective effect, it will either increase in frequency (beneficial mutations) or decrease in frequency (detrimental mutations). The case of a single beneficial mutation has already been studied in the previous chapter. Here I detail two cases of interaction between selection and mutation. First, a detrimental mutation that recurrently appears at the same locus will lead to cryptopolymorphism. Second, because traits under selection generally have a polygenic basis, I show how Fisher's geometric model helped to think about the distribution of fitness effects.

1.3.4.1 Cryptopolymorphism

Let's consider a biallelic locus A, with two alleles, A in frequency $p(t)$ and marginal fitness effect w_A and a in frequency $q(t) = 1 - p(t)$, w_a, so that A mutates to a with probability μ and a mutates to A with probability ν. After one generation of selection and mutation, we have:

$$p(t+1) = p(t)(1-\mu)\frac{w_A}{\overline{w}} + (1-p(t))\nu\frac{w_a}{\overline{w}} \tag{1.141}$$

$$p(t+1) = p(t)(1-\mu-\nu)\frac{w_A}{\overline{w}} + \nu \tag{1.142}$$

Haldane (1927) studied the diploid case with dominance, so that genotypes AA have fitness 1, genotypes Aa have a fitness $1-hs$, and genotypes aa have a fitness $1-s$, were s is the selection coefficient as define above, and h is a dominance coefficient. We can then look at a mutation selection balance and solve $\Delta p = p(t+1) - p(t) = 0$. By making the assumption that $\nu \ll \mu$, *i.e.* that the probability of reversion mutation is negligible, we can find the generic equilibrium frequency of a deleterious allele a:

$$q^* = \frac{(1+\mu)hs - \sqrt{h^2s^2(1+\mu)^2 - 4(1-2h)\mu s}}{2(2h-1)s} \tag{1.143}$$

This simplifies greatly if we suppose that $s \gg \mu$ and $h \gg \sqrt{\frac{\mu}{s}}$:

$$q^* \approx \frac{\mu}{hs} \tag{1.144}$$

In the particular case where $h = 1$, *i.e.* a dominant:

$$q^* = \frac{\mu}{s} \tag{1.145}$$

and $h = 0$, *i.e.* A dominant:

$$q^* = \sqrt{\frac{\mu}{s}} \tag{1.146}$$

This last case corresponds to the situation known as cryptopolymorphism, where a deleterious mutation can be maintained in a population at low frequency.

1.3.4.2 Fisher geometric model

In *The Genetical Theory of Natural Selection*, R.A. Fisher (1930), synthesized his vision of adaptation in a geometrical model. Fisher's Geometric Model (FGM) can be interpreted as an integration of the infinitesimal model in a multidimensional phenotypic landscape. It is mostly used to study the role of mutational effect in the adaptive response thanks to the small number of parameters it incorporates and the implicit inclusion of epistasis. For a review of this approach, see Tenaillon (2014).

An individual is represented by a point in a Euclidian multidimensional phenotypic space, where the axes correspond to combination of traits. The number of independent axes, or dimensionality n, represents the phenotypic complexity of an organism. An optimum is defined for all traits so that all individuals close to it are under stabilizing selection. Fisher used an isotropic model meaning that fitness decays similarly along all axes, but more complicated fitness functions can be included. Lande and Arnold

(1983) for example used the following quadratic decay function. $w(d) = \mathrm{e}^{-\frac{d^2}{2}}$ where d is the distance to the optimum. One can extend this function to a more informative one as Tenaillon et al. (2007):

$$w(d) = \mathrm{e}^{-\alpha d^Q} \tag{1.147}$$

where:

- d is the phenotypic distance to the optimum of the reference genotype
- α and Q are robustness and epistasis parameters influencing the decay rate and the curvature of the fitness function.

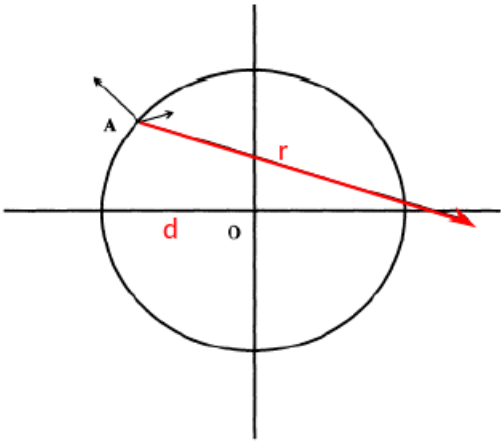

Figure 1.16: **Fisher's geometric model.** The model is represented here in a two-dimensional space ($n = 2$). Any individual A at distance d from the fitness optimum O has the same fitness. A mutation changes the position of the individual in the landscape and can be represented by it's vectorial effect **r** (in red). The illustration clearly shows that most mutations are expected to be detrimental, especially when their effect on fitness is large.

Mutations are included in this model through a vector that translates the parental phenotype to a new position in the phenotypic space. For the sake of simplicity, it is often assumed that a mutation has no preferred direction and affects all phenotypic axes, in what is called universal pleiotropy. Epistasis arises from the curvature of the fitness function because a mutation of vectorial effect **r** will not have the same effect on fitness if it arises in an individual far or close from the optimum. In other words, the effect of a mutation depends on the parental genetic background.

Fisher and later Hartl and Taubes (1996) derived from the model the insightful fraction of beneficial mutations P_{ben}. For $n > 10$ and the scaled distance to the optimum $x = \frac{r}{2d}\sqrt{n}$:

$$P_{ben}(x) \approx \sqrt{\frac{1}{2\pi}} \int_x^\infty \mathrm{e}^{-\frac{u^2}{2}}\, du = \frac{1}{2}\,\mathrm{erfc}(\frac{x}{\sqrt{2}}) \tag{1.148}$$

With erfc the complementary error function, $P_{ben}(x) < \frac{1}{2}$. Therefore, we find the intuitive result that most mutation are deleterious (Figure 1.16). Furthermore, when mutation effects are small compared to the distance to the optimum $\frac{r}{d} << 1$, $P_{ben}(x)$ tends to one half. This is in agreement with both the micro-evolutionist point of view that suggests that adaptation proceeds through infinitesimal steps towards an optimal value, and the quantitative genetics point of view (infinitesimal model). Another interesting result concerns pleiotropy, measured through the dimension of the fitness landscape. As pleiotropy increases, n increases and $x = \frac{r}{2d}\sqrt{n}$ decreases. Hence, as the pleiotropy of the mutation increases, its probability to be beneficial decreases.

In Fisher Geometric Model, most mutations tend to be deleterious and the smaller their effects, the higher their probability to be advantageous.

Overall, this model integrates a rich model of mutation or epistatic interactions that emerges from few simple parameters: a fitness function, a genotype to phenotype map integrating pleiotropy, and a well defined mutation model. However, Fisher's geometric model failed to recognize the important role of genetic drift that a mutation encounters and has to overcome to go to fixation. Further refinements have been made: see (Orr, 1998, 2006; Martin and Lenormand, 2008), at the cost of extra approximations concerning mutations. An important contribution was to tackle the dynamics of fixation of beneficial mutations where evolution consists in the successive fixation of beneficial mutations. As the distance to the optimum gets smaller and smaller, the average effect of mutations that ultimately get fixed decreases at an exponential pace (Orr (1998)). The whole distribution of mutations effects (see Figure 1.7) can also be computed from this model, but see Tenaillon (2014) and Martin and Lenormand (2006).

We will note that Fisher's geometric model primarily focuses on the role of *de novo* mutations with no explicit standing variation and is widely used to model adaptation of asexual populations, while the infinitesimal model focuses more on the adaptive process of sexual populations with some standing genetic variation. "In that process, adaptation is not a single genotype moving in the phenotypic space, but rather a population of recombining and mutating genotypes, which can be characterized in most conditions by a Gaussian cloud." (Tenaillon, 2014)

1.3.5 Interaction between drift and selection

As already discussed before, random genetic drift can counteract the effects of selection and may results in the fixation of a detrimental mutation or the loss of a beneficial mutation by chance. Again, both population and quantitative geneticists derived mathematical models to describe and predict the interaction between drift and selection. One one hand, the diffusion equations provide with the probability of fixation of a single mutation at one locus in a finite population. On the other hand, statistical models were developed in quantitative genetics to predict changes of genetic variance through time.

1.3.5.1 Fixation probabilities and the diffusion equations approximation

Let's focus on a Wright-Fisher haploid population of finite size N, and one diallelic locus, with to allele A and a such that allele a confers a fitness effect of $1+s$, with s the selection coefficient. We already know that the frequency of a beneficial allele should increase, while that of a deleterious allele should decrease. However, the stochasticity of genetic drift can cause this trend to dramatically change, even causing the frequency of a beneficial allele to decrease and the frequency of a deleterious allele to increase. We mentioned in paragraph 1.2.4.1, that selection has a smaller effect when A is either very common or very rare. We can wonder what is the probability of fixation of allele a in a finite population, with competing drift and selection. Kimura (1955), proposed an approximative approach called the diffusion equation approximation. This method has become a standard technique of population genetics. The diffusion equation approach uses a Markov chain model similar to what we use in paragraph 1.2.1. However, the solution to our question requires an extensive knowledge of stochastic processes that are beyond the scope of this introduction. Briefly, it requires the approximation of a discrete time Markov chain to a continuous time Markov process, for which the mathematical properties of an operator, called infinitesimal generator, are used to derive a stochastic differential equation. The diffusion approximation is accurate when selection is weak compared to genetic drift, *i.e.*: $2Ns >> 1$ for haploid population.

If we call u the probability that allele a, in initial frequency p, will get fixed, we can solve the aforesaid stochastic equation and show that:

$$u(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \tag{1.149}$$

If we consider a *de novo* mutation in frequency $\frac{1}{N}$, we have:

$$u(\frac{1}{N}) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \tag{1.150}$$

A graphical representation is given figure 1.17.

From figure 1.17 and equation 1.150, we can see that the probability for a new mutation to go to fixation is almost independent of selection for small population size. With the number of individuals increasing, the efficiency of selection increases, as the greater the selection coefficient, the greater the fixation probability. However, note that for an increasing population size, the fixation probability decreases. In other words selection acts on the convexity of the probability function, and can be seen as a sieve orienting the Brownian motion.

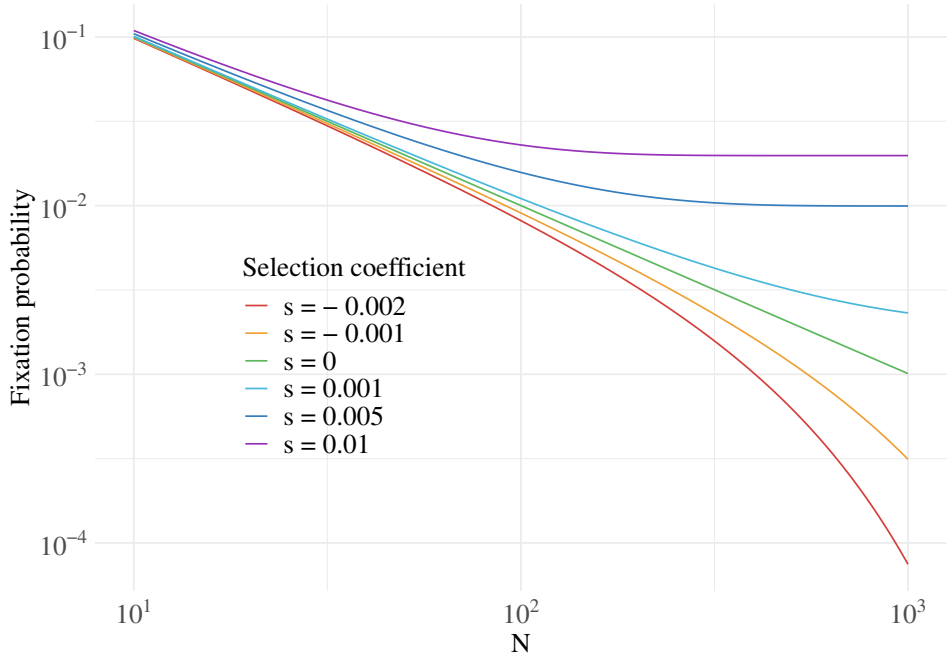

Figure 1.17: **Fixation probability of a new mutation according to the population size**

The limits of application of these equations concern spatially structured populations as sub-population sizes are typically small. Houchmandzadeh and Vallade (2010) have proposed an alternative to the diffusion equation in population genetics, through the use of probability generating function to derive partial differential equations. They use this technique to derive analytical results for the Moran process with selection, among others.

Note that these models do not consider a lot of parameters that are highly influential on the adaptive process, as population size bottlenecks, geographical structure of populations, lineage-specific mutation rates, strong selection and environmental variation. Furthermore these equations describe only the behavior of one locus, and multilocus complications, like hitchhiking of neutral variation with selected alleles at linked loci, are not described.

1.3.5.2 The quantitative genetics framework

If we adopt the vantage point of quantitative genetics and consider a trait without much focus on the underlying loci, we can try to have a glimpse at the evolution of the phenotypic values of a finite population under selection. A traditional approach, that is widely used, notably by breeders, is the animal model, whose name comes from its first implementation in animal breeding that heavily relies on the use of pedigrees, to account for drift and selection. Our "Mixed Models Reminders" gives a quick overview on the mathematical foundation of Best Linear Unbiased Predictor (BLUP) analysis, allowing us to estimate the breeding values of individuals. Several implementation of this method have been produced and are used in breeding programs, or natural population (Kruuk, 2004). We will further exemplify the "animal model" and the "gametic model" presented in great details in Lynch et al. (1998).

1.3.5.2.1 Mixed Models Reminders

In statistics, mixed models are derivatives from the standard Gaussian linear model, where some observations $Y = Y_1, ...Y_n$ are described as the addition of a *fixed* effect that depends on explanatory variables, and a random error following a Gaussian $\mathcal{N}(0, \sigma^2)$ distribution. Errors ϵ_i are supposed to be independent ($\text{Cov}(\epsilon_i, \epsilon_j) = 0$) and to follow the same distribution (homoscedasticity). The linear model writes:

$$Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n) \tag{1.151}$$

with Y the vector of observations of dimensions n, X the fixed design matrix for factors or the fixed matrix containing the explanatory variables for quantitative ones, θ containing the fixed unknown parameters to estimates, I_n the identity matrix and σ^2 the residual variance. For example, with three observations $\{Y_{11}, Y_{12}, Y_{21}\}$, coming from two different populations with means μ_1 and μ_2 respectively, we have $\theta' = \{\mu_1, \mu_2\}$, and

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.152}$$

Contrary to the simple linear model, the mixed model makes it possible to distinguish different sources of variability and to account for correlations between observations, *e.g.* genetic correlations in our case. Therefore the mixed model distinguishes itself from the standard Gaussian linear model based on assumptions of independence and homoscedasticity. Hence, we move to a model that can be written as:

$$Y \sim \mathcal{N}_n(X\theta, \Sigma) \tag{1.153}$$

where Σ is of dimension $n \times n$ and accounts for covariances between observations.

The expected value of the response vector Y keeps the same shape $X\theta$ than before. Note that by construction Σ is symmetrical of dimension $n \times n$, hence $\frac{n(n+1)}{2}$ parameters, that can't be free, because only n observations are possible. Therefore, Σ must have a certain structure governed by a number of parameters called "variance parameters", contained in a vector ψ. The parameters of the model are therefore on one side θ for expectation and ψ for variance. This limitation in the number of possible parameters variance can be both seen as a disadvantage, as the mixed model doesn't allow to directly *estimate* the levels of random factors (only to *predict* them), but has the advantage to better estimate the variance parameters.

Example: common garden experiment For example, let's consider samples from m different populations of the same species, numbered $i = 1, 2, ...m$, each represented by n_i unrelated individuals, numbered $j = 1, 2, ...n_i$ randomly sampled in the population. We will suppose that the populations evolved independently from a common ancestor. The model writes

$$Y_{ij} = \mu + U_i + E_{ij} \tag{1.154}$$

where U_i represents the population effect. All U_i and E_{ij} are supposed normally distributed, centered and independent:

$$\{U_i\} \text{ i.i.d., } U_i \sim \mathcal{N}(0, \gamma^2) \text{ and } \{E_{ij}\} \text{ i.i.d., } E_{ij} \sim \mathcal{N}(0, \mathbb{V}(E)) \tag{1.155}$$

By construction, we have $\mathbb{E}(Y_{ij}) = \mu$ and $\mathbb{V}(Y_{ij}) = \gamma^2 + \mathbb{V}(E)$. This model has three parameters: the mean $\theta = \mu$ and two variance components:

$$\psi = \begin{bmatrix} \gamma^2 \\ \mathbb{V}(E) \end{bmatrix} \tag{1.156}$$

where γ^2 represents the variance between populations. However, individuals from the same population are not independent, because they share a common genetic background or a common *provenance* effect. We have:

$$\text{Cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \gamma^2 & \text{if } i = i' \\ 0 & \text{if not.} \end{cases} \tag{1.157}$$

We can rewrite this model in a matrix form as:

$$Y = X\theta + ZU + E \tag{1.158}$$

where Z is a design matrix $(n \times m)$ constructed here as follow:

$$Z_{a,i} = \begin{cases} 1 & \text{if individual} \quad a = (i,j) \quad \text{comes from population } i, \\ 0 & \text{if not.} \end{cases} \tag{1.159}$$

U is a random vector (of dimension $m \times 1$) such that $G = \mathbb{V}(U) = \gamma^2 I_m$ is the population variance-covariance matrix. We find

$$\mathbb{E}(Y) = \mathbb{E}(X\theta + ZU + E) = X\theta \tag{1.160}$$

and,

$$\Sigma = \mathbb{V}(Y) = \mathbb{V}(X\theta + ZU + E) = Z\,\mathbb{V}(U)Z' + \mathbb{V}(E) = ZGZ' + \mathbb{V}(E) = \gamma^2 ZZ' + \mathbb{V}(E) \tag{1.161}$$

Here, Σ is a block diagonal matrix defined as:

$$\Sigma = \begin{bmatrix} R & 0 & ... & 0 \\ 0 & ... & ... & ... \\ ... & ... & ... & 0 \\ 0 & ... & 0 & R \end{bmatrix} \text{ with } R = \begin{bmatrix} \gamma^2 + \mathbb{V}(E) & \gamma^2 & ... & \gamma^2 \\ \gamma^2 & ... & ... & ... \\ ... & ... & ... & \gamma^2 \\ \gamma^2 & ... & \gamma^2 & \gamma^2 + \mathbb{V}(E) \end{bmatrix} \tag{1.162}$$

It follows the laws of Y conditional on U and the joint distribution of both vectors:

$$Y|U \sim \mathcal{N}_n(X\theta + ZU, \sigma^2 I_n) \tag{1.163}$$

$$\begin{bmatrix} Y \\ U \end{bmatrix} \sim \mathcal{N}_n(\begin{bmatrix} X\theta \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & ZG \\ GZ' & G \end{bmatrix}) \tag{1.164}$$

Using the properties of Gaussian vectors, on can write the conditional laws of U knowing Y:

$$U|Y \sim \mathcal{N}_n(GZ'\Sigma^{-1}(Y - X\theta), G - GZ'\Sigma^{-1}ZG) \tag{1.165}$$

BLUP:
In practice, θ and ψ can be estimating through several methods. Maximum-Likelihood and Restricted-Maximum-Likelihood (*REML*) are the most used. We can then define the predictor (the term *predictor* is used for random effect, instead of *estimator* reserved for fixed effects, like θ):

$$\hat{U} = GZ'\Sigma^{-1}(Y - X\hat{\theta}) \tag{1.166}$$

C. R. Handerson (1963) showed, that **knowing** ψ, this estimator is the *Best Linear Unbiased Predictor* or BLUP. And proposed an easiest way to inverse Σ, reducing dramatically the complexity of the computations.

The *Best Linear Unbiased Estimators* (BLUEs) of θ, is simply the least-squares estimators:

$$\hat{\theta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \tag{1.167}$$

1.3.5.2.2 The animal model

The animal model is a mixed model that incorporates the laws of inheritance and takes into account pedigree relationships between measured individuals. In it's simplest form, it consists in the direct observation y_i of n individuals, indexed with $i = 1..n$. The observation can be decomposed as in (1.75):

$$y_i = \mu + a_i + e_i \tag{1.168}$$

were μ is the fixed factor corresponding to the population mean, and $a_i = 1/2a_{si} + 1/2a_{di}$ is the additive effect, also called *breeding value* of individual i, that depends on the additive effects transmitted by its father (sire si) and mother (dam di). All other effects, including non-additive effects and environment fall into the residual term e_i. The model can be rewritten in a matrix form as any other mixed model: or

$$Y = X\theta + ZU + E \text{ with } X = \begin{bmatrix} 1 \\ ... \\ 1 \end{bmatrix}, \theta = \mu, Z = I_n, U = \begin{bmatrix} a_1 \\ ... \\ a_n \end{bmatrix} \tag{1.169}$$

The prediction equations can be derived as previously. However, we need to take into account that two individuals may be related, *i.e* share a common ancestor. The great force of the animal model is therefore to relax the hypothesis of independence between animals. Indeed, if i and j belong to a same finite population, then

$$\text{Cov}(Y_i, Y_j) = \text{Cov}(a_i, a_j) = 2\varphi_A^{ij}\,\mathbb{V}(A) \tag{1.170}$$

where φ_A^{ij} is the *coefficient of coancestry*, *i.e.* the probability that an allele randomly drawn in i is identical by descent to a randomly drawn alleles from j. Then, the genetic variance covariance matrix G can also be written $G = \mathbb{V}(U) = \mathbb{V}(A) \times A$, where A is the $n \times n$ relatedness matrix that contains the coancestry coefficients, and $\mathbb{V}(A)$ is the additive component of the genetic variance.

The notion of *identity by descent* was formally introduced by Malécot (1948) and is illustrated in Figure 1.18. It describes the probability that two genes derive without mutation from a same ancestral gene and takes into account mendelian segregation: the probability of transmission from parent to offspring is 1/2.

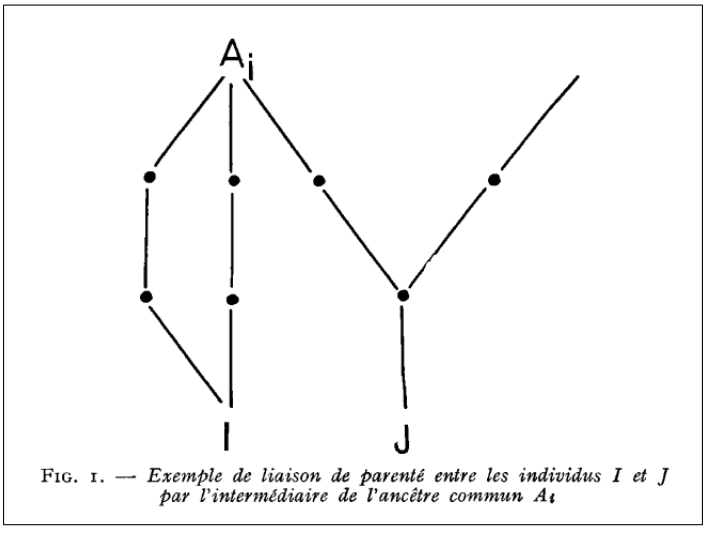

FIG. 1. — *Exemple de liaison de parenté entre les individus I et J par l'intermédiaire de l'ancêtre commun Ai*

Figure 1.18: Illustration from G. Malecot (1972)

Then, the coancestry coefficient can be computed from any known pedigree (Figure 1.18) as

$$\varphi_A^{I,J} = \sum_k \left(\frac{1}{2}\right)^{(nI_k + nJ_k)} \left(\frac{1 + f_{A_i}}{2}\right) \tag{1.171}$$

where k is the number of lineages that relate I and J through the common ancestor A_i, and nI_k (nJ_k) the number of generations between I (J) and A_i. f_{A_i} is the inbreeding coefficient of the ancestor, *i.e.* the probability that the two alleles of the ancestor are identical by descent.

The animal model allows to estimate the additive genetic variance $\mathbb{V}(A)$ in a finite population from the measurement of the phenotypic values, provided the pedigree of the measured individuals is known. It also allows to **predict** the genetic value (breeding value) of any individual from the population knowing its pedigree relationships with the observed individuals.

BLUP-animal model can be applied to natural populations (Kruuk, 2004). The model has the advantage of explicitly accounting for both, the effects of random genetic drift and selection, that merely result in changing the pedigree. However the model does not take into account the introduction of new variation through mutations. It implicitly assumes that all genetic variation is additive and come from ancestral polymorphisms. It's wide application for long-term survey of the evolution of phenotypic variation has been recently criticized (Hadfield et al., 2010).

Extension of the animal model Notice that the model can be extended to the case of dominance, by introducing the coefficient $\varphi_D(i,j)$ of double coancestry : the probability that the two alleles of i are identical by descent to the two alleles of j. In that case, we have:

$$\mathrm{Cov}(Y_i, Y_j) = 2\varphi_A^{ij}\,\mathbb{V}(A) + \varphi_D^{i,j}\,\mathbb{V}(D) \tag{1.172}$$

The table below gives usual values for covariances between related individuals in panmictic populations:

Relatedness	φ_A	φ_D	$cov(Y_i, Y_j)$
Parent-Offspring (PO)	1/4	0	$1/2\,\mathbb{V}(A)$
Full-Sibs	1/4	1/4	$1/2\,\mathbb{V}(A) + 1/4\,\mathbb{V}(D)$
Half-Sibs	1/8	0	$1/4\,\mathbb{V}(A)$
Individual with himself	1/2	1	$\mathbb{V}(A) + \mathbb{V}(D)$

Notice that if dominance exists but is not explicitly introduced in the model, dominance terms will be incorporated in the residual in (1.168). In that case, any departure from random mating will result in $\mathrm{Cov}(e_i, e_j) \neq 0$ and will bias the estimations. This hypothesis might also be violated through environmental correlations of phenotypic values.

1.3.5.2.3 The gametic model

For some situations like animal breeding, one might be interested to know the breeding values of let's say bulls, for characters that are expressed only through their female progenies, like milk production. We can use the gametic model to model the additive genetic value of each offspring in terms of its parents breeding values. Let's call a_{si} and a_{di} the breeding values of the sire and the dam of individual i, so that:

$$a_i = \frac{1}{2}(a_{si} + a_{di}) + e_{ai} \tag{1.173}$$

where e_{ai} correspond the meiosis effects W. Thus we can write the previous animal model as:

$$y_i = \mu + a_i + e_i = \mu + \frac{1}{2}(a_{si} + a_{di}) + (e_{ai} + e_i) \tag{1.174}$$

The same assumptions works in the gametic model as in the previous animal model for the residual e_i. However, one need to take into account here the segregation errors. For non-inbred parents, $\mathbb{V}(e_{ai}) = \frac{1}{2}\,\mathbb{V}(A)$, and for inbred parents, this generalizes to $\mathbb{V}(e_{ai}) = \frac{1}{2}(1 - \frac{F_{si}+F_{di}}{2})\,\mathbb{V}(A)$, F_{si} and F_{di} being the inbreeding coefficients of the sire and dam of i.

1.3.6 Interaction between selection and recombination

As in the previous previous demonstration of the breeder's equation, let's consider a trait from the vantage of the infinitesimal model. We have seen that selection does not change the additive variance if we suppose infinitesimal allelic effects and linkage equilibrium. However, for loci with strong allelic effects, *i.e.* traditional population genetics models show that we can expect changes in allele frequencies impacting the additive variance. Bulmer (1971) demonstrated the existence of an other effect: the fact that selection creates a (negative) correlation between loci, or linkage disequilibrium, therefore impacting (reducing) the additive genetic variance and the subsequent selection response until an equilibrium is reached. Epistasis is another factor that contributes to changes in the additive variance under selection.

1.3.6.1 Bulmer's Effect: linkage disequilibrium generated by selection, and subsequent reduction in additive variance

Let's consider the effect of selection in a polygenic additive model, were the breeding value of individual i is the sum of the alleles brought by the two parents:

$$g_i = a_{si} + a_{di} = \sum_k (a^k_{si} + a^k_{di}) \tag{1.175}$$

The additive genetic variance writes

$$\mathbb{V}(A) = \sum_k \left(\mathbb{V}(a^k_{si}) + \mathbb{V}(a^k_{di}) + \mathrm{Cov}(a^k_{si}, a^k_{di})\right) + 2\sum_k \sum_{l>k} \left(\mathrm{Cov}(a^k_{si}, a^l_{si}) + \mathrm{Cov}(a^k_{si}, a^l_{di}) + \mathrm{Cov}(a^k_{di}, a^k_{di})\right) \tag{1.176}$$

with $\sum_k \left(\mathbb{V}(a^k_{si})\right) = \sum_k \left(\mathbb{V}(a^k_{si})\right) = \mathbb{V}(a)$ being the **genic** variance. Under random mating, the covarianceCov(a^k_{si}, a^k_{di}) vanishes. The other covariance term accounts for linkage disequilibrium between loci. Hence, the additive variance can also be written as:

$$\mathbb{V}(A) = 2\,\mathbb{V}(a) + d \tag{1.177}$$

where d accounts for covariance between loci due to linkage disequilibrium.

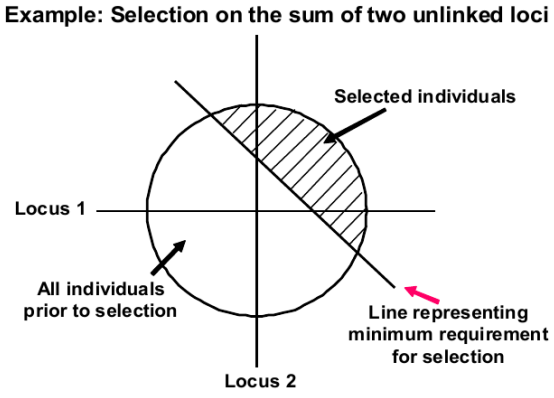

Figure 1.19: **Illustration of the Bulmer effect in a two loci case.** Selection acts on the sum of the effects at two loci and creates a negative covariance between their allele frequencies in the population of selected individuals

In the case of truncation selection (see Figure 1.11) that acts on the sum of individual effects at several loci, it can be easily shown that selection will create a negative linkage disequilibrium between selected parents (Figure 1.19).

Indeed, Bulmer (1971) showed that the expected change in the variance during a single generation of selection is equal to

$$\mathbb{V}(Y_{t+1}) - \mathbb{V}(Y_t) = \frac{h_t^4}{2}(\mathbb{V}(Y_{s,t}) - \mathbb{V}(Y_t)) = d_{t+1} \tag{1.178}$$

where $\mathbb{V}(Y_t)$ is the phenotypic variance of the population at generation t, and $\mathbb{V}(Y_{s,t})$ the phenotypic variance among selected individuals. Because d_{t+1} is negative, truncation selection leads to the decrease of the additive genetic variance from one generation to the next:

$$\mathbb{V}(A)_{t+1} = 2\,\mathbb{V}(a)_t + d_{t+1} \tag{1.179}$$

This effect is not negligible, even under the infinitesimal model (Van Grevenhof et al., 2012). However, recall that linkage disequilibrium is broken-down by recombination. Hence, an equilibrium is expected for the additive variance, between the reduction due to Bulmer's effect and the increase due to the occurrence of new gametic combinations through recombination.

1.3.6.2 Epistasis role in the response to selection and transient response

Epistasis is defined as this ubiquitous biological phenomenon that arises when the phenotypic effect of an allele depends on its genetic context. For example, let's consider a trait determined by two biallelic loci A/a and B/b with the following genetic values :

genotype	AA	Aa	aa
BB	$a+b+abw$	b	$-a+b-abw$
Bb	a	0	$-a$
bb	$a-b-abw$	$-b$	$-a+-b+abw$

with a, b and w being positive integers, and w is an epistatic coefficient that measures the strength of the interaction between the two loci. In this model, changing a A allele by a a allele has a negative effect for BB homozygotes, and a positive effect for bb homozygotes.

In a panmictic population, with p_A the frequency of allele A, and $q_A = 1 - p_A$ the frequency of allele a, the mean of individuals homozytous for the B locus is $\mu_{BB} = b + (p_A - q_A)a(1 + bw)$, and the average effect of allele A is:

$$\alpha_A^{BB} = p_A\,(p_A(a + b + abw) + q_a(b) - mu_{BB}) = q_A a(1 + bw) \tag{1.180}$$

Similarly, the average effect of allele A among bb homozygotes is $\alpha_A^{bb} = q_A a(1 - bw)$. Altogether, if p_B (q_B) is the frequency of allele B (b), we have:

$$\alpha_A = p_B^2\alpha_A^{BB} + 2p_B q_B \alpha_A^{Bb} + q_B^2\alpha_A^{bb} = q_A\,(a + (p_B - q_B)bw) \tag{1.181}$$

Therefore, in presence of epistasis ($w \neq 0$), additive effects of individual loci do not only depend on the frequency of alleles at the target locus, but also on allele frequencies at the other loci, event though they are not polymorphic.

Moreover, additive effects are not sufficient to explain the genetic value G_{ijkl} of any of the nine different genotypes. Neglecting interactions between loci, we have:

$$G_{ijkl} = \mu_G + (\alpha_i + \alpha_j + \delta_{ij}) + (\alpha_k + \alpha_l + \delta_{kl}) + \epsilon_{ijkl} \tag{1.182}$$

where α are the additive effects of alleles, and δ the dominance effects. From this equation, we can see that interaction between loci can arise from different class of interactions: additive $\times$ additive ($\alpha\alpha$), additive $\times$ dominance ($\alpha\delta$), dominance $\times$ dominance ($\delta\delta$). But this kind of decomposition expands exponentially as the number of loci increases (take for example $\alpha\alpha\alpha$, $\alpha\alpha\delta$, $\alpha\delta\delta$, $\delta\delta\delta$ for three loci). Hence we can write the decomposition of the genetic variance of a population as:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + ... \tag{1.183}$$

Epistasis is an universal phenomenon that has two consequences on the genetic variance components in a population:

- The additive genetic variance depends on the effects of all the loci that determine the trait, not only the polymorphic ones Hill et al. (2008). Any change in allele frequencies could result in modifications of the adaptive landscape.
- Additional variance components need to be taken into account to predict parent-offspring relationship and the response to selection.

As compared to additive models, three effects of epistasis are documented:

- *Epistasis increases the unpredictability of the response to selection* by increasing the variance of the variance of additive effects (Dillmann and Foulley, 1998). This results in changes in the response to selection due to stochastic changes in the additive variance.
- *Epistasis transiently inflates the response to selection* through the increase of the *effective* additive variance. Indeed, tacking into account non-additive effects like additive by additive interactions, but also correlation between environments, the parent-offspring regression can be written as:

$$b_{f/o} = \frac{h^2}{2} + \frac{1}{\sigma^2_P}\left(\frac{\sigma^2_{AA}}{4} + \frac{\sigma^2_{AAA}}{8} + ... + \text{Cov}(E_p, E_o)\right) \tag{1.184}$$

Assuming a linear bi-parental regression gives the expected response to selection after one generation of selection:

$$R = 2b_{f/o}S = h^2 S + \frac{S}{\sigma^2_P}\left(\frac{\sigma^2_{AA}}{2} + \frac{\sigma^2_{AAA}}{4} + ... + \text{Cov}(E_p, E_o)\right) \tag{1.185}$$

By considering only additive×additive epistasis, we have:

$$R = S\left(h^2 + \frac{\sigma^2_{AA}}{2\sigma^2_P}\right) \tag{1.186}$$

Griffing (1960) shows, that if we relax selection for t generation, the permanent response to selection will be less that the previous transient response inflated by epistasis, as recombination will break loci associations:

$$R = S\left(h^2 + (1-c)^t \frac{\sigma^2_{AA}}{2\sigma^2_P}\right) \tag{1.187}$$

Hence, even with the transient role of epistasis to the selection response, few generations of panmixia will make the selection response converge to the classical breeder equation.

- *Epistasis constantly inflates the response to selection* because at each generation, phenotypes represent only a limited part of possible variation, given the range of possible multiloci combinations. Even infinitesimally small changes in allele frequencies trigger the emergence new genetic combinations (Barton, 2017). Figure 1.20 shows the mean response to selection and corresponding changes in additive (VA) and additive by additive (VA) variance under the infinitesimal model. Simulation shows that with epistasis, the response to selection can be much more important, albeit a long-term decrease of additive genetic variance comparable to the neutral expectation.

1.3.7 Interaction between selection, drift, and mutation

Population genomics focuses on, first recovering signatures of selection in the genome, and subsequently ask which trait were affected by selection. Quantitative genomics instead seeks at understanding the selection response for a trait and then asks what are the implication on the genomes. In real life, mutations affect a set of phenotypic traits that are "then" submitted to selection. I will briefly present here a quantitative genetics extension of the mixed model equations that accounts for mutations. I will then present population genetic models that trace the genomic signatures of beneficial mutations that increase in frequency in a population. Finally, I will review recent extensions on polygenic adaptation.

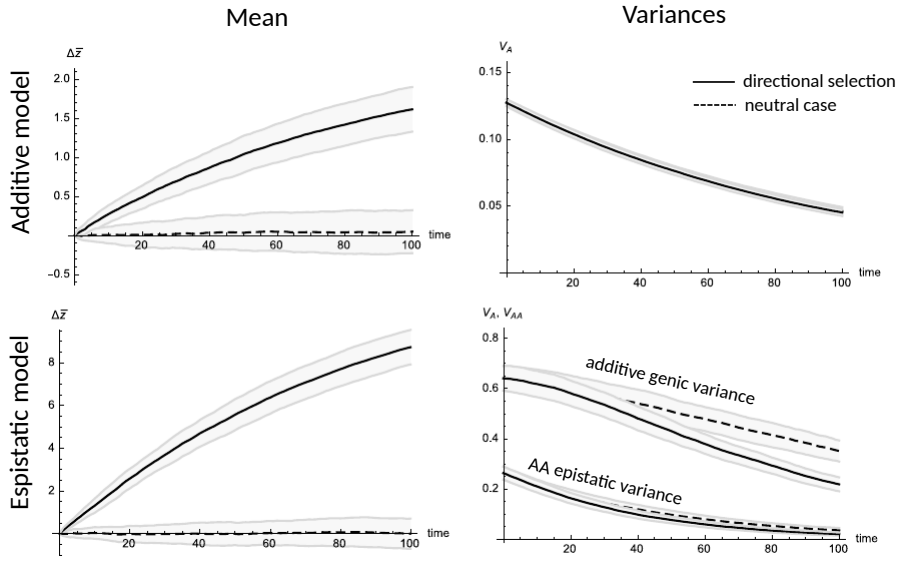

Figure 1.20: **Response to selection in the infinitesimal model.** Directional selection (lines) is compared with the neutral case (dotted lines) with (bottom) or without (top) epistasis. The model considers $M = 1000$ loci in a population of $N = 100$ haploid individuals. At each locus, alleles have the same affect but an arbitrary sign$\gamma = +/- \frac{1}{\sqrt{M}}$. Epistasis was simulated by drawing a small fraction $1/M$ of interactions between pairs of loci w_{ij} in a normal law of null expectation and standard deviation $\frac{4}{\sqrt{M}}$, with possibly $w_{ij} \neq w_{ji}$. Phenotypic trait value is defined as $z = \delta\gamma + \delta^T w \delta$, with $\delta = +/- 1/2$. Initial allele frequencies were drawn in a U-shaped Gamma distribution with mean $\bar{p} = 0.2$ and variance $0.2\bar{p}\bar{q}$. Parents selected to form the next generation are drawn accordingly to their selective value with $W = e^{\beta z}$. Grey areas represent $+/-$ one standard deviation around the mean. Adapted from Barton et al. (2017).

1.3.7.1 The BLUPM: how to take into account *de novo* mutation

In the previous section, the mixed model used to estimate breeding values ignored the effects of mutations, which is reasonable in many settings. However, for multiple generation pedigrees like the ones found in long-term evolution experiments, mutations cannot be ignored. Wray (1990) proposed a model that aims at considering *de novo* mutation, based on the idea that a new mutation introduces additional dependencies between breeding values in the progenies of a mutated progenitor. The author showed:

$$\mathbb{V}(U) = \sigma^2_{A0} \left(A + \frac{\sigma^2_m}{\sigma^2_{A0}} \sum_{g=1}^{G} A_g \right) \tag{1.188}$$

where:

- σ^2_m corresponds to the rate of additional additive mutational variance due to mutations,
- σ^2_{A0} corresponds to the initial additive genetic variance found in the base population due to standing variation,
- A is the standard relationship matrix ignoring mutation,
- A_g is a relationship matrix that ignores ancestors between generations 0 and $g - 1$.

Under this model the A matrix is updated each generation by adding an additional relationship between progenies of the previous generation. This implies, that the rate of input of new mutational variance is constant through generations, which might be relevant for large population size. However, this model fails to account for stochasticity of the mutational process which may not be negligible for small mutation rates and small mutational targets, and when mutations may occur outside the focal generation. Here, the model considers a constant mutational input through time.

1.3.7.2 Genomic signatures of the adaptive processes

In our attempt to exemplify how the interplay between evolutionary forces changes allele frequencies during adaptation, we need to present three major scenarios developed by population geneticists, that predicts distinct genomic footprints.

Hermisson and Pennings (2017) define selective sweeps as the patterns in genomic diversity that are caused by recent adaptation due to genetic hitchicking (Maynard Smith and Haigh, 1974). These footprints arise when beneficial mutations increase due to positive selection, and the genealogical histories of the samples at the surrounding loci are distorted because of genetic draft, *i.e.* the increase in frequency of the surrounding alleles at neighbor loci due of linkage disequilibrium. Recall that positive selection tends to decrease the fixation time compared to the neutral case, hence in a coalescent approach, looking back at time, selection decreases the coalescence time distorting the genealogies to star-shape genealogies.

1.3.7.2.1 Hard sweep

This case can be defined formally as the scenario for which the TMRCA is more recent than the time at which the selective pressure started to act (T_S, the onset time of selection). For example, when a change of environment precedes the appearance of a new beneficial mutation. This leads to a strong decrease in genetic diversity at the selected loci and it surrounding region. Because coalescence is quicker than recombination in this case, recombination events mainly occur on external branches (the leaves). *I.e* most neutral variation at the flanking region is found as singletons, with an increase in both low and high derived variant frequencies compare to the neutral SFS, Figure 1.21 A.

The example of the population genetics behind a hard sweep: (Adapted from Walsh and Lynch (2018))

In order to get a glimpse at the mechanisms acting behind a hard selective sweep, we can consider the following model. First, consider diploid population of size N with a neutral allele m, associated to a biallelic locus A, with 2 alleles A (in frequency $p(t)$) favored by selection and a (in frequency $1-p(t)$) as usual. We can define $q(t)$ the frequency of m at generation t after the onset of selection leading to the sweep. Hence $q(\infty)$ represents the frequency of m after the sweep completion. If there is no recombination between locus A and allele m, it follows that the final frequency of m is $q(\infty) = 1$. However, if there is recombination, we can expect $q(t)$ to be a function of c/s: the ratio of the recombination rate c that tends to separate the neutral allele from the selected site, and s the selection coefficient of allele A at the favored site that drags the allele m.

To study the evolution of the allele frequency of m, we can use the conditional decomposition of $q(t)$:

$$q(t) = q_A(t)p(t) + q_a(t)(1-p(t)) \tag{1.189}$$

so that $q_A(t)$ corresponds to the frequency of allele m on haplotype carrying allele A, and $q_a(t)$ corresponds to the frequency of allele m on haplotype carrying allele a. When allele A is fixed, we have $q(\infty) = q_A(\infty)$ which is the fraction of A-bearing haplotypes still carrying allele m. We can then define $\delta_q(t) = q_A(t) - q_a(t)$ the difference in the frequency of m on the two backgrounds, so that $\delta_q(t) \neq 0$ implies linkage disequilibrium between A and m. We are also interested in $\Delta_q = q(\infty) - q(0)$ the change in allele frequency of m due to the selective sweep.

We can see that if we normalize Δ_q by the initial difference in frequency of m on the two backgrounds $\delta_q(t)$, we can access the strength of the selective sweep:

$$f_s = \frac{\Delta_q}{\delta_q(0)} = \frac{q(\infty) - q(0)}{q_A(0) - q_a(0)} \tag{1.190}$$

Said otherwise, f_s corresponds to the fraction of the initial excess of association between m and A, after the sweep completion.

In the case of a hard sweep, the new selected allele A is introduced in one copy in the population, hence $q_a(0) \approx q(0)$, as all, minus one copy of m, are on haplotypes carrying allele a. Following the same argument, $q_A(0) = 1$.

Therefore, we can rewrite f_s as:

$$f_s = \frac{\Delta_q}{\delta_q(0)} = \frac{q(\infty) - q(0)}{q_A(0) - q_a(0)} = \frac{q(\infty) - q(0)}{1 - q(0)} \tag{1.191}$$

$$\Longleftrightarrow$$

$$q(\infty) = q(0) + \Delta_q = q(0) + f_s\delta_q(0) = q(0) + f_s(1 - q(0)) = f_s + q(0)(1 - f_s) \tag{1.192}$$

Hence the final frequency of m can be expressed as a function of its initial value and of the sweep strength. Furthermore, we can note that the frequency of the neutral allele m in both backgrounds only changes trough recombination, so that we have:

$$\delta_q(t) = q_A(t) - q_a(t) = (1 - c)^t(q_A(0) - q_a(0)) \approx \delta_q(0)\,\mathrm{e}^{-ct}$$

Barton (2000) showed that we can express Δ_q as:

$$\Delta_q \approx \delta_q(0)p(0)^{c/s} \tag{1.193}$$

Hence, when $c/s \ll 1$ *i.e.* selection is much stronger than the recombination rate as in the vicinity of the selected allele, most initial linkage remains but this tend to 0 as more distant site are considered (*i.e.* c/s increases).

Using this result, we can estimate the strength of a hard sweep as a function of c/s:

$$f_s = \frac{\Delta_q}{\delta(q)} = p(0)^{c/s} \tag{1.194}$$

And this allows accessing the frequency of m after the sweep completion, as a function of its initial value, the frequency of allele A, the recombination rate and the strength of selection:

$$q(\infty) = p(0)^{c/s} + q(0)(1 - p(0)^{c/s}) \tag{1.195}$$

Again, the strength of the selective sweep increases as c/s diminishes, either because selection increases or because c decreases.

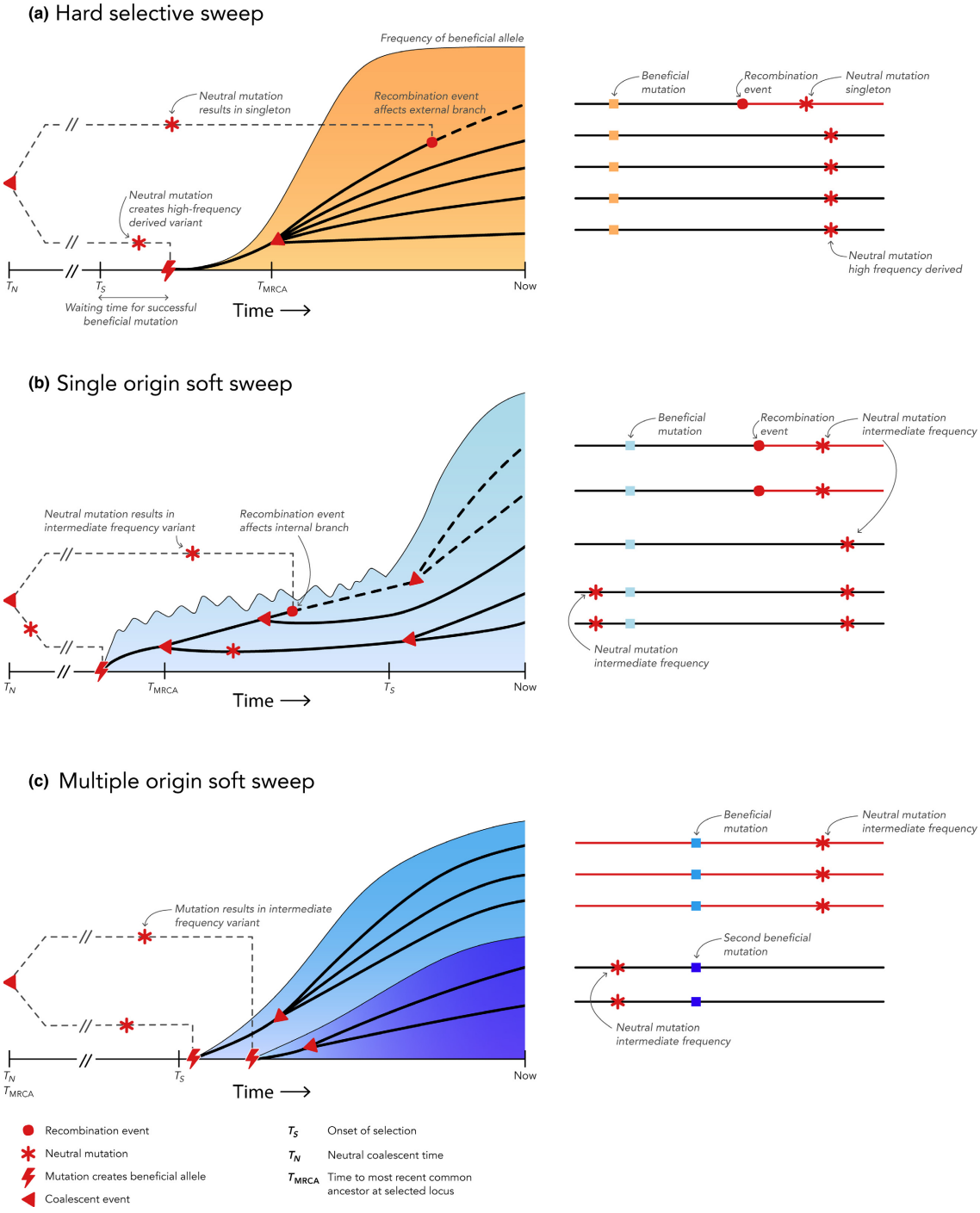

Figure 1.21: Hard and soft sweep types. Colored regions mark the frequency of copies of the beneficial allele that still have descendants at the time of sampling. Black and dashed lines show coalescent histories at linked sites. On the right, mutation and recombination events are also shown on haplotypes of the five sampled individuals. (a) For a hard sweep, the time to the most recent common ancestor (TMRCA) at the selected site T_{MRCA} is more recent than the onset of selection T_S. All ancestral variation at tightly linked sites is eliminated. Recombination leads to low-frequency and high-frequency derived variants in flanking regions. (b) For a single-origin soft sweep from standing genetic variation, TMRCA occurs before the onset of selection so that multiple haplotypes carrying the beneficial allele are present in the 'standing phase' before T_S. Early recombination produces variants at intermediate frequencies. (c) The beneficial allele traces back to multiple origins. Each origin introduces an ancestral haplotype, typically at intermediate frequency. Hermisson and Pennings (2017)

1.3.7.2.2 Soft sweep

Here, with T_N, the expected neutral coalescent time, we have $T_N > T_{MRCA} > T_S$ (looking backward in time). This happens for example when an environmental change occurs and selection acts upon pre-existing standing variation. Hence, the MRCA predating the onset of selection, recombination had more time to create new haplotypes.

- If the same beneficial mutation have been selected from several haplotypes, this scenario is referred to as a 'single origin soft sweep'.
- If several beneficial mutations have been selected from several haplotypes, one can talk of 'multiple origin soft sweep'.

Figure 1.21 is extracted from Hermisson and Pennings (2017) and illustrates the link between sampled haplotypes on the right and their genealogical history on the left, in the case of hard sweep, soft sweep originating from a single origin, and soft sweep originating from multiple origins.

Effective population size primarily determines the likelihood of soft sweeps. Hence, Messer and Petrov (2013) have shown that when θ (four times the product of effective population size and the beneficial mutation rate) is equal or above 1, and selection is strong enough, adaptation proceeds from multiple de novo mutations or standing variation. Below 1, soft sweeps' contribution diminishes with theta.

1.3.7.2.3 Polygenic adaptation: Towards an integration of population and quantitative genetics

We have seen that the genomic signatures left by a selective sweep is a function of the selection intensity, that is significantly diminished under soft sweep scenarios. Hence, weak selection using standing variation at a number of loci is the worst-case scenario for sweep detection. In the most extreme case of the infinitesimal model, we have seen that a significant shift in the mean can occur with almost no detectable allele-frequency changes. Hence, in a case where a trait is encoded by many loci of small effects, we expect polygenic adaptation to leave almost no genomic signal.

Chevin and Hospital (2008) extended the theory of hitchhiking of a locus affecting a quantitative trait to account for background genetic variation at other loci. The dynamics at the focal quantitative trait locus is related to the initial genotypic value and to the background genetic variance of the trait. Using analytical approaches, they derived the conditions for a selective sweep, and the expected patterns of genetic diversity diversity at the end of the sweep. They showed that selective sweeps were possible, even for highly polygenic traits. However, while the dynamics of the focal locus clearly depends on the mode of selection (linear, exponential or Gaussian), the selection coefficient of the beneficial mutation decreases in time because of background genetic variance (Figure 1.22). Overall, phenotypic traits exhibiting clear-cut molecular signatures of selection may represent only a small subset of all adaptive traits.

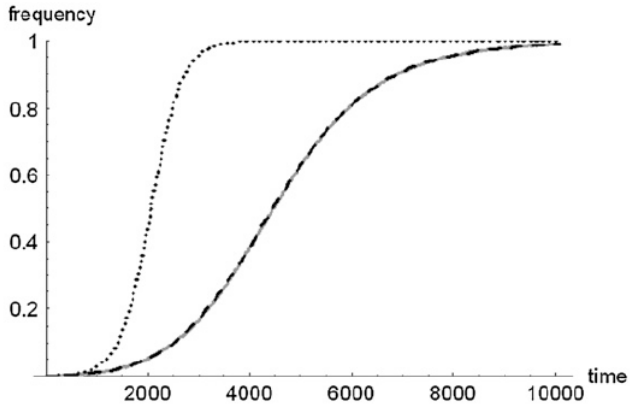

Figure 1.22: Dynamics of fixation of a beneficial mutation for a polygenic trait submitted to directional selection. The dotted line represents the dynamics without background genetic variance. With background genetic variance, the selection coefficient of the beneficial mutation decreases through time therefore increasing the time to fixation (adapted from Chevin and Hospital (2008)).

Stetter et al. (2018) used extensive forward-time simulations and machine-learning algorithms to understand the complex interplay of mutation, selection, and demography on the genomic signatures of polygenic adaptation. In agreement with Chevin and Hospital (2008), the authors showed that selective sweeps occur even for traits under relatively weak selection where the genetic background explains most of the variation. Most sweeps occur from variation segregating in the ancestral population, nevertheless new mutations can be important for traits under strong stabilizing selection that undergo a large optimum shift. They emphasized the role of bottlenecks and expansions on the overall genetic variation as well as the relative importance of sweeps from standing variation and the speed with which adaptation can occur.

Höllinger et al. (2019) used newly devised analytical framework based on Yule branching processes and computer simulations to show the major role of the population-scaled background mutation rate θ_{bg}. For a focal locus, θ_{bg} measures the mutation rate at all redundant loci in its genetic background that offer alternative ways for adaptation. To do so, they modeled a binary polygenic trait (such as resistance to a pathogen) with negative epistasis. Compared to the infinitesimal model, the authors emphasized that for a finite number of loci, polygenic adaptation pattern are expected to occur if alleles are able to hamper the rise of alleles at other loci via negative epistasis for fitness. Indeed, they stated that one would otherwise only observe parallel sweeps. Note that negative fitness epistasis is frequently found in empirical studies, *e.g.* due to Michaelis-Menten enzyme kinetics, and is implicit to the Gaussian selection scheme. Recall that we have seen previously in Fisher geometric model, that a mutation might overshoot the optimum depending on the genetic background. Höllinger et al. (2019) divided their model of the adaptive process into two phases, a stochastic one and a deterministic one. They derived for several extensions of their model (like linkage disequilibrium, partial redundancy...) the joint distribution of allele frequency ratios of minor over major loci after the stochastic phase. Then, they showed that the ratio is given by an inverted Dirichlet distribution, from which we can extract the distribution of allele frequencies. Furthermore, they predicted that the distribution of these ratios is preserved during the deterministic phase. Interestingly, they showed that adaptation proceeds by sweeps for small $\theta_{bg} \leq 0.1$ and that small polygenic allele frequency shifts require large $\theta_{bg} \geq 100$, in compliance with the infinitesimal model. In the large intermediate regime, they observed a heterogeneous pattern of partial sweeps at several interacting loci.

1.3.8 Long term Selection Response and the limits to selection

A pervasive question in quantitative genetics concerns the limits to the response to selection. While living forms are in constant evolution, what should happen in a finite population submitted to a constant selection in a constant environment ? On one hand, there could be of course physiological limits for some trait values. On the other hand, mutational targets may not be infinite and drift may limit genetic variation.

1.3.8.1 Selection Limits under Drift

Drift impact on the additive genetic variance:

We have previously shown that the allele diversity decreases through time as a function of the population size, ($\mathbb{E}(H_t) = H_0(1 - \frac{1}{N})^t$, for a haploid population). As the genic variance is a function of allele frequencies that change because of drift, we can also show that for a diploid population, we have:

$$\sigma_a^2(t) = \sigma_a^2(0)\left(1 - \frac{1}{2N_e}\right)^t \tag{1.196}$$

Hence, the evolution of the additive variance of quantitative traits submitted to drift alone is expected to decay geometrically trough time as a function of the effective population size.

Robertson's Limit of cumulative selection response:

To show the evolution of the selection response through time, Robertson (1960) ignored the effect of linkage disequilibrium, assumed $\sigma_A^2(t) = \sigma_a^2(t)$, a constant phenotypic variance, and a constant effective

population size. He showed:

$$R(t) = i\frac{\sigma_A^2(t)}{\sigma_P} = \left(1 - \frac{1}{2N_e}\right)^t i\frac{\sigma_A^2(0)}{\sigma_P} = \left(1 - \frac{1}{2N_e}\right)^t R(1) \tag{1.197}$$

Hence, we can derive the expected cumulative response to selection by first noting that:

$$\sum_{t=0}^{T}\left(1 - \frac{1}{2N_e}\right)^t \approx 2N_e\left(1 - \mathrm{e}^{-\frac{T}{2N_e}}\right) \tag{1.198}$$

Therefore the cumulative response at generation T is equal to:

$$R^{(T)} = 2N_e\left(1 - \mathrm{e}^{-\frac{T}{2N_e}}\right)R^{(0)} \tag{1.199}$$

As T tends to infinity, the upper limit for the expected cumulative response due to the exhaustion of genetic variance by drift is $2N_e$ times the initial response.

According to Walsh (2010), this upper limit is a reasonable one as soon as the selection intensity or the effect of drift remains small at any given locus.

Chevalet's extension for finite population size and loci number:

Based on (Lande, 2008) and Felsenstein (1977) work, Chevalet (1994) extended Bulmer's equation to the case of finite population size (N_e) and loci number (n) experiencing truncation selection under an infinite allele model.

$$\Delta\sigma_a^2(t) = -\left[\frac{\sigma_a^2(t)}{2N_e} + \left(1 - \frac{1}{N_e}\right)\frac{kh_{ss}^2\sigma_A^2(t)}{2n}\right] \tag{1.200}$$

$$\Delta d(t) = -\frac{1}{2}\left[\left(1 + \frac{1}{N_e}\right)d(t) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{N_e}\right)kh_{ss}^2\sigma_A^2(t)\right] \tag{1.201}$$

$k = i(i - z_{[1-p]})$ with $z_{[p]}$ defined as $\Pr(U \leq z_{[p]}) = p$ and U a unit normal distribution.

Taking into account a finite number of loci and linkage disequilibrium, the limits to selection response scales as the inverse of both effective population size and number of loci.

1.3.8.2 Selection Limits under Drift and Mutations

If we consider the additive variance due to mutations at one generation, σ_m^2, we can write the following recursion equation:

$$\sigma_A^2(t) = \left(1 - \frac{1}{2N_e}\right)\sigma_A^2(t-1) + \sigma_m^2 \tag{1.202}$$

This recursion equation can be solved to give approximately:

$$\sigma_A^2(t) \approx 2N_e\sigma_m^2 + \left(\sigma_A^2(0) - 2N_e\sigma_m^2\right)\mathrm{e}^{-\frac{t}{2N_e}} \tag{1.203}$$

Therefore, one can compute the expected genetic additive variance at drift mutation equilibrium, when t tends to infinity.

$$\sigma_A^2(\infty) = 2N_e\sigma_m^2 \tag{1.204}$$

Following Hill (1982a); Wei et al. (1996); Weber and Diggins (1990) the expected cumulative response after T generations can be computed as:

$$R^{(T)} \approx 2N_e \frac{i}{\sigma_P} \left[t\sigma_m^2 + \left(1 - \mathrm{e}^{-\frac{t}{2N_e}}\right) \left(\sigma_A^2(0) - 2N_e\sigma_m^2\right) \right] \tag{1.205}$$

When $t \to \infty$, the term $t\sigma_m^2$ corresponds to the asymptotic response and dominates as the initial standing genetic variance $\sigma_A^2(0)$ vanishes.

The balance between the exhaustion of initial genetic variance by drift and the constant input of new mutations leads to a constant response that scales with the mutational variance σ_m^2.

1.4 The role of new mutations in the response to selection : Saclay's Divergent Selection Experiments (SDSE)

Experimental evolution is a method of choice for testing predictions in evolutionary biology. It consists in monitoring populations that evolve under controlled conditions. The selection pressure can be known (*e.g* selection to increase or decrease the value of a trait) or determined by the controlled environmental conditions (*e.g* Lenski's experiment on the adaptation of a *E. Coli* strain to lab conditions). In all cases, the dynamics of adaptation can be monitored at the genomic scale and at the traits level.

Divergent selection experiments (DSEs) consist in splitting-up an initial population into at least two sub-populations submitted to divergent selection regimes.

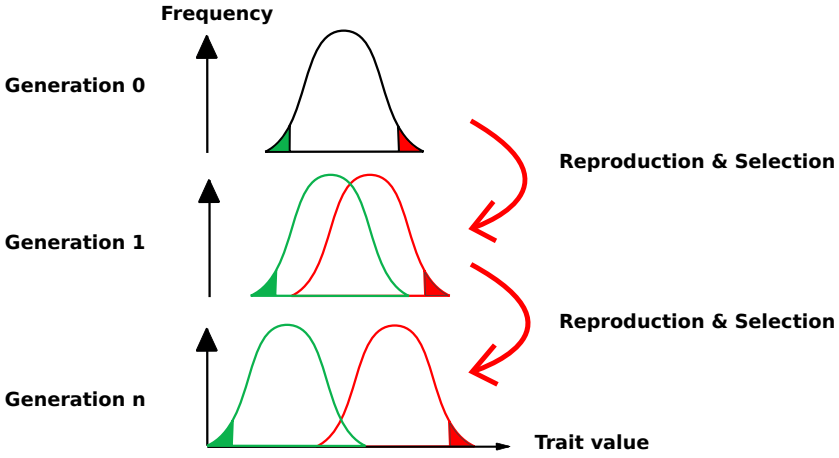

Figure 1.23: **Divergent Selection Experiments principle** Starting from an initial population represented here by its distribution of phenotypic values, extreme individuals (with the lowest and the highest phenotypic value) are selected, and define two new independent populations that are selected and reproduced independently at each new generation.

Saclay's DSEs were set-up in 1993 from maize inbred lines seed-lots characterized by a very small amount of standing genetic variation (Durand et al., 2010). Populations were selected for early or late flowering date, which is an important agronomical trait for maize cultivation with a polygenic basis. Although maize is an allogamous species, the mating system employed was selfing. The other conditions of the experiment were constrained by practical reasons and led to a high selection pressure (1% of the population each year), high drift (5 individuals selected in each population) but relatively high census number in each population (500 individuals).

The initial questions of the experiments were the following :

- How far can we change flowering time in maize from new mutations ?
- What are the genes targeted by mutations ?
- What is the dynamics of fixation of beneficial mutations ?

Altogether, a constant response to selection was observed in both directions. At the beginning of my PhD, genomic resources had been produced (RNA-Seq) from evolved lines from the 13th generation of selection, and an evaluation trial (generations G0 to G18) was planned.

1.4.1 Saclay's divergent selection experiments for flowering time in Maize

1.4.1.1 Maize flowering time

The term floral transition designates in plants the developmental process through which a plant switch from vegetative growth to reproductive growth. This key developmental stage governs the whole life of the plant and primarily determines flowering time, but also linked characteristics such as the plant height, total number of leaves, or grain fill.

Flowering time has been shown to be a complex trait displaying in maize large ranges of variablity, with flowering dates occurring between 35 to 120 days after sowing (Colasanti and Muszynski, 2009). In maize, Buckler et al. (2009) showed that that the trait is associated to roughly hundred small effects QTLs, characteristic of highly additive polygenic traits.

However, very few genetic determinants have been characterized at the molecular level. The gene network associated to floral transition, represented Fig. 1.24, is characterised by the interaction of several different metabolic pathways integrating environmental and developmental information, such as the gibberellins pathway (*e.g.* with the Dwarf 8 gene region (Thornsberry et al., 2001; Andersen et al., 2005; Camus-Kulandaivelu et al., 2006)), the aging pathway, the circadian clock, the photoperiod pathway (*e.g.* with the gene ZmCCT (Ducrocq et al., 2009; Coles et al., 2010) and CONSTANS-like gene (conz1) (Miller et al., 2008)), or the autonomous pathway (*e.g.* Indeterminate gene (Id1) (Colasanti et al., 1998)). These different pathways are integrated through the interaction of different genes, such as the activator ZMM4, the vegetative to generative transition 1 gene (vgt1) (Salvi et al., 2007), or ZCN8 (Meng et al., 2011) produced in the leaf navigating to the meristem to interact with the delayed flowering1 gene (dlf1) (Muszynski et al., 2006).

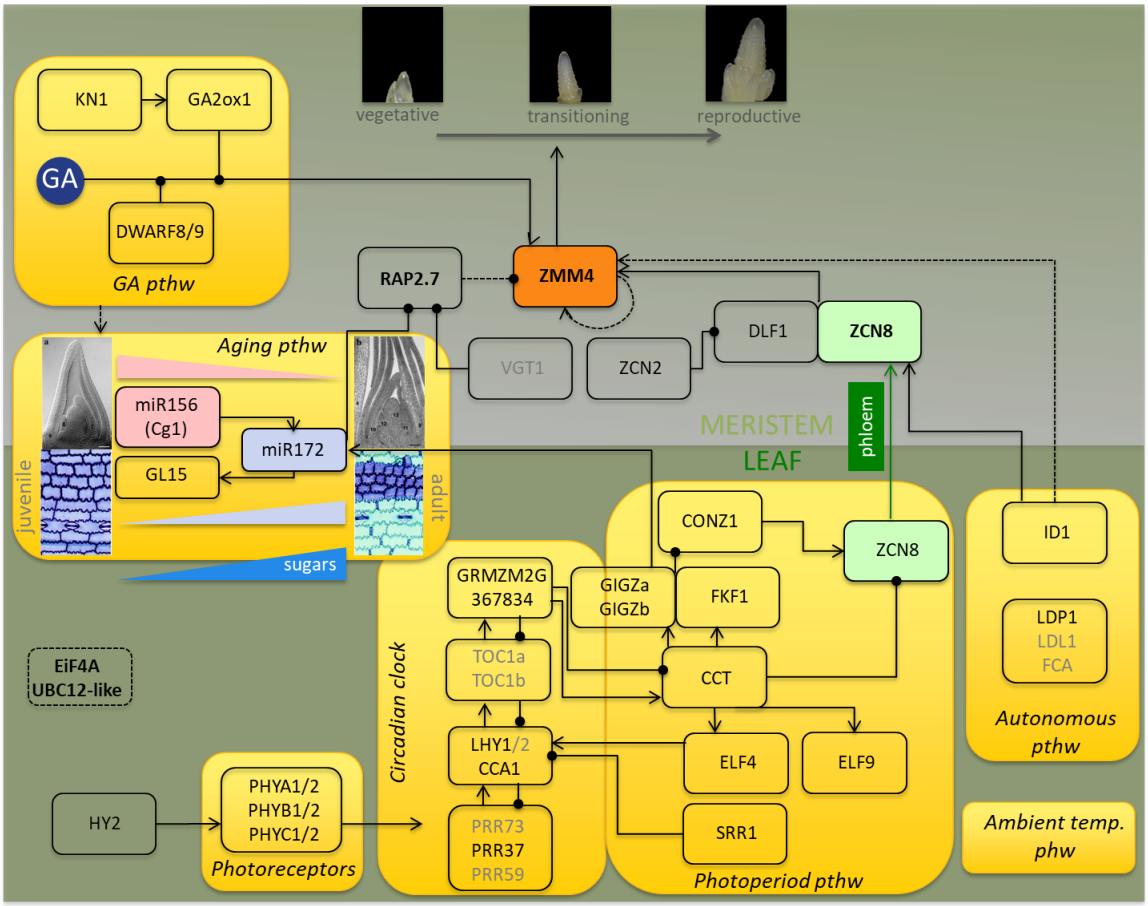

Figure 1.24: **Flowering time known underlying gene network** Extracted from Tenaillon et al. (2018)

1.4.1.2 Initial inbred lines

Two certified maize inbred lines were used as initial populations: an american flint, F252, registered in 1979 by the company Agri-Obtention, and a late iodent dent, MBS847 (MBS for later on), registered by the company Mike Brayton Seeds in 1982 (Durand et al., 2010).

We do not exactly know the selection scheme used to produce these two inbred lines. However, we can make the assumption that they have been produced through a classical maize breeding scheme developed by Hopkins in 1908. This method is called the ear to row method adapted to produce inbred lines by Single Seed Descent, where inbreds are obtained from F1 hybrids after 6 to 8 generation of selfing and selection. During this step, a number of plants are selected on the basis of their phenotype. They are selfed and seeds are harvested on single plant basis. A single row of 10 to 50 full-sibling progeny coming from a same ear is raised. The progeny rows are evaluated and the best progenies for the chosen traits are identified. It corresponds to a Single Seed Descent scheme where the whole ear is planted to allow a better evaluation of the genotype of the parent. Selfing is done manually to control for outcrossing. The last generation consists in producing a pre-base seed lot by harvesting the selfing progenies of the selected individual. The pre-base seed lot is then submitted to control for homogeneity before registration. Then the base seed lots are produced by two generations of bulk in isolation. In France, registration of a variety requires three main criteria that are homogeneity, stability and distinction. This process is controlled and certified by SOC http://www.gnis.fr/. It ensures a level of homozygosity for the base seed lots.

Our experiment started in 1993, long after the first registration of these varieties. This suggests

that these base seed lots had undergone several generations of multiplication from the initial pre-base population, with the same protocol as the production of base seed lots from pre-base. Therefore, we expect the mean residual heterozygosity to range between $(\frac{1}{2})^8 = 0.39\%$ and $(\frac{1}{2})^{12} = 0.0244\%$ validated in Durand et al. (2015). We also expect polymorphisms between pre-base individuals. This would also lead to fixed differences in the base population. Each inbred line was treated separately as an independent biological replicate of the selection experiment. However, they do not constitute true replicates as different genetic features might differentiate them and lead to different response to selection.

1.4.1.3 Experimental protocol

These long-lasting field experiments have been taking place since 1993 in Gif-sur-Yvette on Saclay's Plateau. All selected seeds are stored at 6°C in a cool chamber, and the pedigree information have been carefully recorded, producing a unique genetic material with corresponding phenotypic information. Furthermore, few changes in the selection scheme have been made throughout the years and are included in the presentation hereafter and schematized Fig. 1.25.

1993 - G_0: For each genetic background, F252 and MBS, 60 plants were grown in fields condition, female flowering time was recorded and the 3 earliest (and 3 latest respectively) plants were selfed and their kernels harvested, constituting the G0 parents of three families of the Early, respectively Late population.

1997 - G_1: 100 progenies of the 3 selected plants per population were grown in a randomized block design, with 25 plants per row sown together at a density of 25000 plants/ha and a spacing between rows of 80 cm. Selfing was made when both male and female flowering occurred, and the selfing date date was recorded in days after July 1st. This step imposes a selection pressure against long Anthesis-Silking Intervals (ASI). The basal kernels were harvested and weighted. We selected in the Early (respectively Late) populations the 10 earliest (resp. latest) flowering plants with the highest kernel weight. This additional selective pressure assured us a lesser inbreeding depression that could further impede our selection process.

1998 and later - G_2 **to** G_n: To better control for environmental effect, each of the 4 rows representing the 100 progenies of a selfed plant were sown in a randomized block design, independently for the populations of MBS and F252. The sowing density was 25000 plants/ha and 80 cm between rows. More accurately, from each selected parents, 100 seeds were distributed into 4 different blocks (25 seeds in each). Each block was constituted of two plots of 11 rows each (one plot for the Early population, one for the the Late). One random row was dedicated to 25 plants of the initial seed lot as control and the 10 remaining rows were randomly sown with 25 seeds of the 10 parents selected previously. Early and Late plots were alternatively distributed along the blocks. The 3 earliest (resp. latest) flowering plants were selfed within each row (except the border plants). Their basal kernel were harvested, weighted and stored in a cool chamber at +6°C. Later on, a second step of selection was made based on additional constraints:

- 10 earliest (resp. latest) plants are selected in each population,
- between two plants of equivalent earliness, the plant with the highest kernel weight was chosen,
- we did not select more than three plants from the same parent,
- we did not select more than two plants from the same row,
- we maintained at least two lineages per population of the initial parents selected in 1993, called families in the rest of the thesis.

Overall MBS genealogy (Appendix A.) can be decomposed into two independent early flowering families subsequently called MP049 and MP052 forming the early population, and two independent late flowering families MT040 and MT053 forming the late flowering population. In F252, two early independent families also constituted the early population. The late flowering population started as two independent families, the late FT031, and the late FT027. Despite a very strong selection response (Durand et al., 2010, 2015), FT027 did not produce seeds at generation 14 (flowered to late in the season), and could not be maintained further. FT031 was then divided into two subfamilies FT317 and FT318 sharing a common ancestor at G3 (Appendix A.)

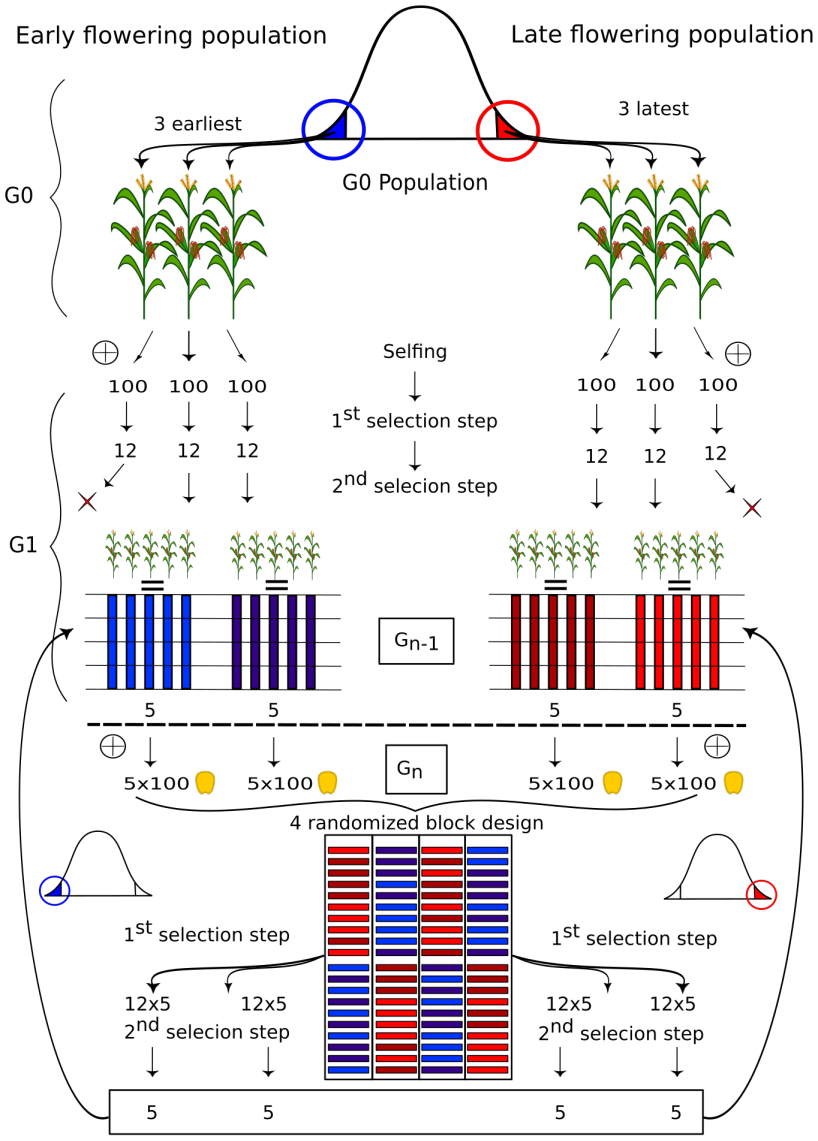

Figure 1.25: **Experimental scheme of Saclay DSEs.** For clarity a single scheme is shown but was replicated for the two DSEs. Starting from an inbred G_0 population with little standing variation ($< 1\%$ residual heterozygosity (Durand et al., 2015)), the three earliest (resp. latest) flowering individuals represented in blue (resp. red) were chosen based on their offspring phenotypic values as the founders of two families forming the early (resp. late) population. For the subsequent generations, 10 (≈ 5 per family) extreme progenitors were selected in a two-step selection scheme among 1000 plants. More specifically, 100 seeds per progenitor were evaluated in a four randomized-block design, *i.e.* 25 seeds per block in a single row. In a first selection step, the $3 \times 4 = 12$ earliest (resp. latest) flowering plants among the 100 plants per progenitor were selected. Then in a second selection step, 10 (≈ 5 per family) individuals were selected within each population based on both flowering time and kernel weight and the additional condition of preserving two progenitors per family from the previous generation.

1.4.1.4 Previously published results

Durand et al. (2010) showed that the response to selection was fast (they focused on the first seven generations) and significant in both the Early-flowering (Early) and the Late-flowering (Late) populations. The observed selection response could not be explained without accounting for mutational input, and estimated values of mutational heritability ranged from 0.013 to 0.025, which corresponds to an upper bound of what is reported in other species. Part of the response was nevertheless attributed to standing genetic variation in the initial seed lot. They remarkably identified one polymorphism initially segregating in the F252 seed lot at a candidate locus for flowering time, which explained 35% of the trait variation within the Late F252 population, more precisely in FT027 (FL1 in Fig. 1.26). Durand et al. (2012), characterized more thoroughly this candidate region and identifying the Eukaryotic Initiation Factor (eIF-4A), and further revealed a high level of sequence and structural variation beyond the 3'-UTR of eIF-4A, including several insertions of truncated transposable elements. They confirmed through association genetics the association of the underlying polymorphism of this gene with flowering time variation and revealed in the association panel the pervasive interactions between allelic variation and the genetic background, pointing to underlying epistasis. They highlighted the importance of pleiotropic effects of the candidate polymorphism on various traits including flowering time, plant height, and leaf number. Finally Durand et al. (2015) refined the results of Durand et al. (2010) by including the first 16 generations to their analysis (Fig. 1.26), and observed striking similarities between the two inbred lines: a plateau is reached after 7 generations in Late families which they interpreted as resulting from physiological limits. They used 42 markers derived from both Methyl-Sensitive Amplification and Amplified Fragment Length Polymorphisms (AFLP) and showed that 13 of them were strongly associated with flowering time variation. Their fast fixation throughout DSEs' pedigrees resulted in strong genetic differentiation between populations and families.

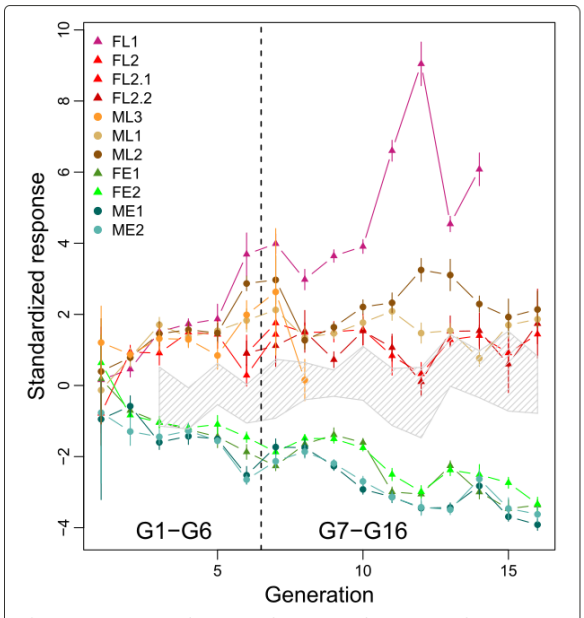

Fig. 2 Response to selection in the *F252* and *MBS* DSEs. The average flowering time of each family as estimated by difference to the first generation per unit of control residual standard deviation is plotted against time in generation. A separate color code for lines and points has been attributed to each family. Vertical lines indicate standard deviation within families. Data from the *F252* and *MBS* DSEs have been treated separately, except for control lines. The shaded area represents the range of flowering time variation in control lines (data available only from generation G3)

Figure 1.26: Observed response to selection in all families, extracted from (Durand et al., 2015).

1.4.2 Objectives of the PhD

Previous results revealed a paradox between the sustainability of the response to selection and the associated dearth of initial polymorphisms in small selfing population submitted to strong selection (1% of selected individuals) and high drift ($N_e \approx 2.5$). This paradox implies the necessity of incoming *de novo* mutations to sustain the observed selection response. To disentangle the relative contribution of standing genetic variation and *de novo* mutations in the selection response, I adopted 3 mains approaches.

In Chapter 2, I relied on forward individual-based simulations, calibrated from the observed selection response to make predictions on the expected dynamics and the distribution of selected fitness effects of *de novo* mutations.

In Chapter 3, I used a previously published RNAseq data set to detect polymorphic SNPs between one (resp. one) early flowering and one late (resp. one late and one very late) flowering progenitor at G_{13} in MBS (resp. F252). I notably developed a simulation approach with C. Dillmann to characterised a simplistic expected null distribution of initial polymorphisms along the genome in our initial seed lots (SSDs), accounting for selfing, drift and linkage disequilibrium. I then genotyped using a Kaspar^{TM} approach (KBioscience's competitive allele-specific PCR amplification of target sequences and endpoint fluorescence genotyping) a subset of 190 progenitors per inbred line for the detected SNPs. I adapted for our SNP markers a unique parsimony and maximum likelihood inference algorithm based on pedigree information originally developed by Durand et al. (2015) for AFLP markers. I inferred the genotypes of all our selected progenitors, which allowed me to confidently characterise *de novo* mutations and standing genetic variation. I further compared the observed allele frequency dynamics considering both standing and *de novo* polymorphisms, to the simulated *de novo* mutations dynamics of Chapter 2.

Finally, in Chapter 4, I contributed to the phenotypic evaluation of a subset of 190 progenitors per inbred line in a two years common garden experiment (in 2018 an 2019, on Saclay Plateau), subsequently called SdpEval. While several traits were recorded, I decided to focus mainly on the continuous traits: flowering time, plant height and length of the leaf just below the upper ear. I analysed the observed selection response, and the correlative changes of the traits, with the aim of deciphering (i) the role of *de novo* mutational variation in the selection response using Wray (1990) BLUPM model, (ii) the impact of genotype-by-environment interactions. I also associated polymorphisms detected in Chapter 3 to the observed phenotypic variation of the three traits.