1.3 Classical Genetics

The study of heredity began long before the molecular structure of DNA was understood. Several thousand years of experience breeding animals and plants led, eventually, to the idea that hereditary characteristics are passed along from parents to offspring in units, which are termed *genes.*

Classical genetics states that some genes are *dominant* and others *recessive.* For example, suppose we have a certain "heredity unit" symbolized as A_1 whose presence in an individual leads to brown eyes. A variant gene, A_2, sometimes appears in the population; individuals carrying it grow up with blue eyes. Humans, among other *diploid* organisms, carry two genes for each trait, which are called *alleles.* According to the classical concept of dominance, having one dominant allele outweighs the presence of a recessive one. Brown eyes turn out to be dominant in humans, so a person with an A_1A_2 mix of alleles has brown irises, just like one whose alleles read A_1A_1. Only an A_2A_2 individual develops blue irises.

1.3.1 Reproduction

The dynamics of a population depends upon births of new individuals, with possibly novel mutations. To maintain a constant population size this must be accompanied by death of members of previous generations. Even without mutations ($\mu_1 = \mu_2 = 0$ in the previous example), reproduction by birth/death introduces stochasticity in the dynamics (say of the proportion x_1 of allele A_1). To emphasize the role of reproduction, in this section we shall initially neglect mutations, and follow changes in a preexisting diversity of alleles in the population.

Hardy-Weinberg equilibrium: Within diploid organisms, sex and *mating* present additional complications, which we shall ignore by adapting a gene-centered perspective. To see why this may be justified in at least some limit, consider an idealized population consisting of very large number of individuals ($N \to \infty$), where diploid organisms mate randomly with no preference for phenotypic or geographic considerations.[3] The initial population is characterized by the proportions x_{11}, x_{12}, and x_{22} of the genotypes A_1A_1, A_1A_2 and A_2A_2, with $x_{11} + x_{12} + x_{22} = 1$. The composition of the next generation is obtained by considering all possible matings and their outcomes. For example, a pairing of two homozygotes A_1A_1 individuals occurs with probability x_{11}^2, and leads to A_1A_1 offspring. However, a mating of A_1A_1 with A_1A_2, with probability $x_{11}x_{12}$ may lead to either an A_1A_1 offspring, or an A_1A_2 offspring. Assuming no selective advantage for either such offspring, each happens with probability of 1/2. Similarly, the pairing of two heterozygotes A_1A_2 may result in A_1A_1, A_1A_2 and A_2A_2 with probabilities of 1/4, 1/2, and 1/4, respectively. Including all 9 (3×3)

[3]The list of assumptions is extensive, including no mutations, migration, selection; and discrete generations in addition to random mating. The following argument also deals with an infinite population of hermaphrodite, although male/female distinction can in principle be dealt with.

pairings, we arrive at

$$\begin{aligned} x'_{11} &= x_{11}^2 + 2 \cdot \frac{x_{11}x_{12}}{2} + \frac{x_{12}^2}{4} = \left(x_{11} + \frac{x_{12}}{2}\right)^2 , \\ x'_{12} &= 2x_{11}x_{22} + 2 \cdot \frac{x_{11}x_{12}}{2} + 2 \cdot \frac{x_{22}x_{12}}{2} + \frac{x_{12}^2}{2} = 2\left(x_{11} + \frac{x_{12}}{2}\right)\left(x_{22} + \frac{x_{12}}{2}\right) , \\ x'_{22} &= x_{22}^2 + 2 \cdot \frac{x_{22}x_{12}}{2} + \frac{x_{12}^2}{4} = \left(x_{22} + \frac{x_{12}}{2}\right)^2 . \end{aligned} \tag{1.29}$$

(Note that pairings of distinct genotypes involve an additional factor of two, from the degeneracy in their order of selection.)

It is easy to check that the above results are completely equivalent to $x'_1 = x_1$ and $x'_2 = x_2$, where $x_1 = x_{11} + x_{12}/2$ and $x_2 = x_{22} + x_{12}/2 = 1 - x_1$ are the the proportions of alleles A_1 and A_2 in the diploid population. (For example, the first equation above can be recast as $x'_{11} = x_1'^2 = x_1^2$.) Thus, within one generation the alleles are mixed by random reproduction such that the proportion of the three possible genotypes merely reflects the proportion of the allele in the entire population. This so-called *Hardy-Weinberg equilibrium* justifies the gene-centered perspective as a theoretical limit. In fact, within a population of finite size N the frequency x_1 is not constant, but will change stochastically due to random reproduction events as discussed next.

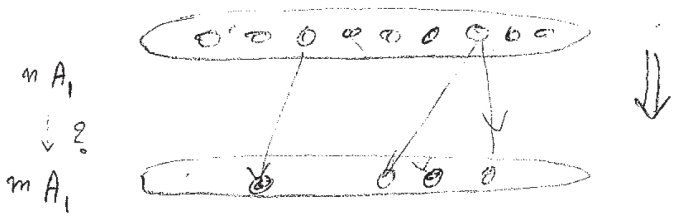
Fisher-Wright (binomial) process: Consider a population with two forms of an allele, say A_1 and A_2 corresponding to blue or brown eye colors. The probability for a spontaneous mutation to occur that changes the allele for eye color is extremely small, and effectively $\mu_1 = \mu_2 = 0$ in Eq. (1.24). Yet the proportions of the two alleles in the population does change from generation to generation. One reason is that some individuals do not reproduce and leave no descendants, while others reproduce many times and have multiple descendants. This is itself a stochastic process and the major source of rapid changes in allele proportions. In principle, this effect also leads to variations in population size. To simplify computations, we initially assume that the size of the population is fixed, and consider the effects of variation later.

Continuing with the gene-centered perspective, consider the following, so called *Fisher-Wright* process starting from the $2N$ alleles in a diploid population of size N. In the model of *binomial selection*, the process or reproduction from one generation to the next is assumed to be as follows: One allele is randomly selected, an exact copy is made for the next generation, while the parent allele is returned to the original pool. This process is repeated $2N$ times to produce the next generation. Let us assume that in the initial population of $2N$ alleles, $N_1 = n = 2Nx_1$ are A_1, and the remaining $2N - n$ are A_2. The population at the next generation may have m individuals with allele A_1, with (transition) probability

$$\Pi_{mn} = \left(\frac{n}{2N}\right)^m \left(1 - \frac{n}{2N}\right)^{2N-m} \left(\begin{array}{c} 2N \\ m \end{array} \right) . \tag{1.30}$$

The process leading to such probability is like reaching into a bag with n balls of blue color and $2N - m$ balls of brown color, recording the color of the selected ball and throwing it

back to the bag. After repeating such selection N times, the probability that the blue color is recorded m times is given by the above binomial distribution. (The probability of getting a blue ball in each trial is simply $n/2N$, and $1 - n/2N$ for brown.) On average, the number of alleles does not change, since $\langle m \rangle = n$ from the binomial distribution (i.e. $\langle x_1' \rangle = x_1$ consistent with Hardy-Weinberg equilibrium). However, there is now a range of possible values of m; clearly the stochasticity arises since some balls can be picked up multiple times (multiple descendants), while some balls are never picked (no offspring). The mathematical consequences of Eq. (1.30) will be explored later on.



1.3.2 Heterozygosity

In the absence of mutations, random reproduction in a finite population inevitably leads to a loss of diversity, known as genetic drift. This loss can be quantified by following the evolution of the *homozygosity* measure, $G = \sum_{i=1}^{k} x_i^2$, where x_i is the proportion of allele i (out of k possibilities) in the population, and its complement, the *heterozygosity* $H = 1 - G$. For the case of $k = 2$, $G = x^2 + (1-x)^2$, and $H = 2x(1-x)$. Let us follow the change in G_t or H_t from generation t to $t+1$ in the Fisher-Wright process. A new homozygote diploid is generated either from duplication of the same chromosome, with probability $1/2N$ for selecting the same chromosome twice, or from two separate chromosomes that share the same allele, resulting in the recursion relations

$$G_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) G_t \,, \quad \text{and} \quad H_{t+1} = H_t \left(1 - \frac{1}{2N}\right) . \tag{1.31}$$

Thus heterozygosity decays exponentially as

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t = H_0 e^{-t/\tau} \,, \tag{1.32}$$

over a time (in generations) scale $\tau = 1/\ln(1 - 1/2N) \approx 2N$.

The above calculating is easily generalized to allow for a population size $N(t)$ varying with generations. The loss of heterozygosity in each generation is related to size of its population, generalizing Eq. (1.32) to

$$H_{t+1} = H_t \left(1 - \frac{1}{2N}\right) = H_0 \prod_{i=1}^{t} (1 - \frac{1}{2N(i)}) \,. \tag{1.33}$$

The long-time decay of H is best captured by considering the logarithm,

$$\ln H_{t+1} \approx \ln H_0 + \sum_{i=1}^{t} \frac{1}{2N(i)} \approx \ln H_0 - \frac{t}{2N_{eff}}\,, \tag{1.34}$$

where

$$N_{eff}^{-1} = \frac{1}{t}\sum_{i=1}^{t} N(i)^{-1}\,. \tag{1.35}$$

If there are large variations in population size, N_{eff} with be close to the smallest value N_{min}. For example, when considering the population of humans of European/Asian descent, there are indications that $N_{eff} \approx 5,000$, presumably the size of the ancestor sub-population that migrated out of Africa around 30,000 years ago.

Mutations are easily included in the above calculation, with the framework of the *infinite allele model*. The latter model assumes that each mutation, at a rate μ leads to a new allele, neglecting the possibility of mutations return mutations. (This is in fact a quite good assumption if the locus under consideration is an entire gene.) The recursion relation for homozygosity in Eq. (1.31) is now simply modified through multiplication by a factor of $(1-\mu)^2$, the requirement that neither of the selected chromosomes has duplicated in reproduction, i.e.

$$G_{t+1} = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) G_t\right](1-\mu)^2\,. \tag{1.36}$$

(Note that in the infinite allele model we do have to worry about appearance of homozygosity due to either back mutation, or double mutations to the same allele.) In corresponding recursion relation for heterozygosity, we shall drop terms of order μ^2 and μ/N, arriving at

$$H_{t+1} = H_t\left(1 - \frac{1}{2N}\right) + 2\mu(1 - H_t)\,. \tag{1.37}$$

The loss of diversity due to genetic drift is now counteracted by mutations; resulting in a net change in heterozygosity of

$$\Delta H_t \equiv H_{t+1} - H_t = -\frac{H_t}{2N} + 2\mu(1 - H_t)\,.$$

The two effects are balanced in *steady-state*, with

$$H^* \approx \frac{4N\mu}{4N\mu + 1}.$$

1.3.3 Selection

We assumed so far that the two alleles are completely equivalent, corresponding to *neutral* evolution. It is likely that one allele is better in the sense of conferring a selective advantage to the individual carrying it. The selective advantage of a genotype is parameterized through

an associated *fitness* that quantifies its number of likely progeny (relative to other genotypes). In our diploid binary allele example, we may associate fitness values of f_{11}, f_{12} and f_{22} to the three genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Indicating the proportion of allele A_1 in the population by $x \equiv x_1 = n/2N$, the average fitness is given by

$$\overline{f}(x) = x^2 f_{11} + 2x(1-x)f_{12} + (1-x)^2 f_{22}\,. \tag{1.38}$$

The expected fractions of off-spring for the three genotypes are thus governed by the relative fitness values of $f_{11}/\overline{f}$, $f_{12}/\overline{f}$ and $f_{22}/\overline{f}$.

After one generation, the frequency x on average changes to

$$\langle x' \rangle = \frac{f_{11}}{\overline{f}} x^2 + \frac{1}{2}\frac{f_{12}}{\overline{f}} \cdot 2x(1-x)\,. \tag{1.39}$$

The expected change in the proportion of the allele is thus given by

$$\begin{aligned}
\Delta x \equiv \langle x' \rangle - x &= \frac{1}{\overline{f}}\left[f_{11}x^2 + f_{12}x(1-x) - \overline{f}x\right] \\
&= \frac{1}{\overline{f}}\left[f_{11}x^2 + f_{12}x(1-x) - f_{11}x^3 - 2f_{12}x^2(1-x) - f_{22}x(1-x)^2\right] \\
&= \frac{1}{\overline{f}}\left[f_{11}x^2(1-x) + f_{12}x(1-x)(1-2x) - f_{22}x(1-x)^2\right] \\
&= \frac{x(1-x)}{\overline{f}}\left[\frac{1}{2}\frac{d\overline{f}(x)}{dx}\right] \\
&= \frac{x(1-x)}{2}\frac{d\ln\overline{f}}{dx}\,.
\end{aligned} \tag{1.40}$$

The above result, known as *Wright's equation* implies that allele frequencies always change so as to maximize the average fitness function $\overline{f}(x)$. A corresponding result holds for a multi-loci situation with a corresponding *fitness landscape* $\overline{f}(x_1, x_2, \cdots, x_n)$.

For ease of computations, in the following sections we shall write the selective advantage for allele A_1 ias

$$\Delta x = \frac{x(1-x)}{2}s\,, \tag{1.41}$$

typically ignoring any x dependence of s.