21.1 Population Genetics

This is the first of a series of lectures we will hold on population genetics. The focus today will be on fitness and epistasis.

**Definition 21.1.** For an alphabet $\Sigma$, a **genotype** is an element of $\Sigma^n$.

Example 1: For any fixed $n$ and $\Sigma = \{0, 1\}$, an example of a genotype is $00110101\ldots$ or any string of length $n$ consisting of zeros and ones. This example is relevant, as the places considered when inspecting a single strand of a genome are often SNPs (single nucleotide polymorphisms) and they tend to take only two forms. The actual state of a SNP in a specific genome is thus determined as 0 or 1, and any string $s \in \{0, 1\}^n$ represents a specific genome.

Example 2: $\Sigma = \{0, 1, 2\}$. This example corresponds to the case of characterizing both strands of a genome, and thus each place consists of two characters that can be either 0 or 1 (as in Example 1). This is indicated by 0/1/2, indicating the situation on both strands at each place. For example, we may represent a DNA sequence in the following manner:

![Figure 21.1. An example of a genotype for $\Sigma = \{0, 1, 2\}$.

**Definition 21.2.** A **genotype space** $\mathcal{G}$ is a subset of $\Sigma^n$, $\mathcal{G} \subseteq \Sigma^n$. A genotype space can be viewed as a collection of genotypes.
Definition 21.3. A genotope $\Pi_G$ is the convex hull of the vectors in $G$.

Example: For $\Sigma = \{0, 1, 2\}$ there will be different genotopes for different $G$'s:

1) For $G = \{0, 1, 2\}$, $\Pi_G = [0, 2]$, the straight line between the points 0 and 2.

2) For $G = \{0, 1, 2\}^2$, all strings of length 2, we get that $\Pi_G$ is the square with vertices $(0,0), (0,2), (2,0), \text{and } (2,2)$.

3) For $G = \{0, 1, 2\}^3$ we get that $\Pi_G$ is the cube with vertices $(000, 020, 200, 220, \ldots, 222)$.

Definition 21.4. A population is a point in the simplex $\Delta_G$.

A population can be viewed as a set of probabilities, indicating the relative amount of each genotype in the population.

Example: For $G = \{0, 1\}^2 = \{00, 01, 10, 11\}$, we have

$$\Delta_G = \{ (p_{00}, p_{01}, p_{10}, p_{11}) \mid \Sigma p_i = 1, p_i \geq 0 \}.$$ 

A population is any point $(p_{00}, p_{01}, p_{10}, p_{11}) \in \Delta_G$.

We can define a mapping $\rho : \Delta_G \to \Pi_G$ specifying a relation between a population of $G$ and its genotope.

For example, in the diploid case we define $\rho(p_{00}, p_{01}, p_{10}, p_{11}) = (p_{10} + p_{11}, p_{01} + p_{11})$.

Mapping back with $\rho^{-1}(v)$ gives for the given vector $v$ all populations that have allele frequencies (the frequencies of observing a 1 at place $i$) as specified in $v$.

Definition 21.5. A fitness landscape is a function $\omega : G \to \mathbb{R}$. We define $\omega_g$ as the value of $\omega$ for genotype $g$.

A fitness landscape associates to each $g \in G$ a real number $w_g$. When considering fitness, one measures the reproductive ability of a genotype, $g \in G$, under selection.

We define penetrance as the probability of getting a disease.

The vector space of all fitness landscapes is the $|G|$-dimensional vector space $\mathbb{R}^G$.

Example**: We can visualize the genotope and a fitness landscape for $G = \{00, 01, 10, 11\}$ as shown in Figure 21.2, where the fitness landscape consists of $\omega_{00}, \omega_{01}, \omega_{10}, \omega_{11} \in \mathbb{R}$, and we can view the height of each $\omega_i$ as the fitness of the corresponding genotype.

We define the epistasis in this case as the linear form $\omega_{00} + \omega_{11} - \omega_{01} - \omega_{10}$. Biologically, epistasis is defined as the interaction between genes. The linear equation characterizes the
change in fitness for the different possible combinations of genes. This can also be viewed from an economic perspective as characterizing the value of a whole ($\omega_{00} + \omega_{11}$) compared to the value of the sum of its parts ($\omega_{01} + \omega_{10}$).

Each fitness landscape has one of the following three types:

1) Additive Landscapes: $\omega_{00} + \omega_{11} - \omega_{01} - \omega_{10} = 0$. In this instance we can pass a plane through the four points.

2) Positive Landscapes: $\omega_{00} + \omega_{11} - \omega_{01} - \omega_{10} > 0$.

3) Negative Landscapes: $\omega_{00} + \omega_{11} - \omega_{01} - \omega_{10} < 0$.

We define $L_G \subseteq \mathbb{R}^G$ in the following way:

$L_G = \{ \omega \mid \text{there exists a linear function defined on } \Pi_G \text{ that agrees with } \omega \text{ at the vertices} \}$.

We define the interaction space, $I_G$, to consist of all linear forms in the unknowns $\omega_g$ that vanish on the subspace $L_G$.

In Example** $\mathbb{R}^G$ is 4-dimensional, $L_G$ is 3-dimensional (as a plane can be determined by 3 points), and $I_G$ is the 1-dimensional space spanned by $\omega_{00} + \omega_{11} - \omega_{01} - \omega_{10}$.

Lemma 21.6. $\dim(I_G) = |G| - \dim(\Pi_G) - 1$

Definition 21.7. A **circuit** is a linear form which has minimal support with respect to inclusion.
The interaction space is spanned by circuits.

Example: For $\mathcal{G} = \{000, 110, 011, 100, 101, 111\}$, we have $|\mathcal{G}| = 6$ and $\text{dim}(I_\mathcal{G}) = 2$. In this example there are four circuits:

$$
\begin{align*}
\omega_{100} - \omega_{101} - \omega_{110} + \omega_{111} \\
\omega_{000} - \omega_{011} - \omega_{100} + \omega_{111} \\
\omega_{011} + \omega_{101} + \omega_{110} - \omega_{000} - 2\omega_{111} \\
\omega_{000} + \omega_{101} + \omega_{110} - \omega_{011} - 2\omega_{100}.
\end{align*}
$$

Typically, the number of circuits is greater than $\text{dim}(I_\mathcal{G})$. However, examining all circuits is of interest, as we see in the following example.

Example: ($\Pi_\mathcal{G}$ is a 3-dimensional cube)

The case of the 3-D cube has the following 20 circuits:

- 6 corresponding to the six faces of the cube. These circuits measure the conditional epistasis between two loci when the third locus is fixed (we refer to place $i$ in $g \in \mathcal{G}$ as locus $i$).
- 6 squares formed by vertices of the cube that slice it into two triangular prisms, which relate the marginal epistasis of two pairs of loci.
- 8 bi-pyramids, which relate the 3-way interaction to the total 2-way epistasis.

In this example four linear forms form a natural basis for the interaction space:

$$
\begin{align*}
u_{110} &= (\omega_{000} + \omega_{001}) + (\omega_{110} + \omega_{111}) - (\omega_{010} + \omega_{011}) - (\omega_{100} + \omega_{101}) \\
u_{101} &= (\omega_{000} + \omega_{010}) + (\omega_{101} + \omega_{111}) - (\omega_{001} + \omega_{011}) - (\omega_{100} + \omega_{110}) \\
u_{011} &= (\omega_{000} + \omega_{100}) + (\omega_{011} + \omega_{111}) - (\omega_{001} + \omega_{101}) - (\omega_{010} + \omega_{110}) \\
u_{111} &= (\omega_{000} + \omega_{001} + \omega_{101} + \omega_{110}) - (\omega_{100} + \omega_{010} - \omega_{001} + \omega_{111}).
\end{align*}
$$

The linear form $u_{110}$ marginalizes over locus 3, measuring the marginal epistasis between locus 1 and locus 2. In the same manner, $u_{101}$ and $u_{011}$ marginalize over locus 2 and locus 1 respectively. Finally, $u_{111}$ measures the three-way interaction among the three loci.

The circuits can be written as combinations of the different linear forms defined. For example, the circuits in the second group of circuits listed above can be written as:

$$
\begin{align*}
u_{110} + \nu_{101} \\
u_{110} - \nu_{101} \\
u_{110} + \nu_{011} \\
u_{110} - \nu_{011} \\
u_{101} + \nu_{011} \\
u_{101} - \nu_{011}.
\end{align*}
$$
21.2 Optional Homework

Prove Lemma 21.6: $\dim(I_G) = |G| - \dim(\Pi_G) - 1$.