# 1
# Quantitative Genetics of Human Traits

## 1.1 Anthropometric variation

During the late 1800s, a criminologist working in France, Alphonse Bertillon, and others developed a sophisticated collection of approaches, collectively known as anthropometrics, for measuring physical (morphological) variation among individuals. Most of the traits that Bertillon considered were what we now call continuous traits, based on measurements (e.g., height, weight, etc) as opposed to discrete traits which are based on counts of unambiguous observations (e.g., eye color, etc). The British scientist Francis Galton analysed the patterns of variation of continuous traits, such as height, among relatives (parents and children) in an attempt to discover general rules for the hereditary transmission of trait values from parents to offspring. Part of Galton's motivation for this research was his belief in eugenics – essentially a movement aimed at improving human races by selective breeding (encouraging "desirable" individuals to reproduce with tax incentives, etc) and discouraging "undesirables" from reproducing by sterilization, and other measures. Heritability of a trait is essential for eugenics to be effective in altering its frequency in populations. Eugenics has now been largely discredited – the greatest abuse of eugenics was carried out by the Nazi's who used it to justify the murder of millions.

Although Galton initiated such studies, it was not until after the rediscovery of Mendel's laws in 1900 that the British mathematician and biologist R.A. Fisher developed the first credible explanation for the patterns of inheritance of continuous traits in humans. Fisher's theory can be used to predict correlations of traits among relatives and played a dominant role in studies of human disease genetics and improvements in animal breeding methods during the first half of the twentieth

century. This branch of populations genetics, which predicts patterns of trait variation among relatives without knowing the specific underlying genes influencing the trait is known as "quantitative genetics." Quantitative genetics still plays an important role in studies of so-called complex genetic traits and diseases (those that are caused by the combied effects of many genes and environment). Many common human diseases such as type II diabetes are complex genetic diseases. Virtually all continuously varying human traits, such as height and weight, are complex traits.

## 1.2 Fisher's model

Let $m$ be the measure of a continuous trait (phenotype) such as height in humans. Assume that $L$ genes (each with two alleles) influence the trait. Each gene (locus) is assumed to have a small effect (of similar magnitude) with each copy of an allele $D$ at a locus increasing $m$ by an amount $+a/2$ and each copy of allele $d$ decreasing $m$ by $-a/2$. The trait value of an individual is

$$m = \sum_{i=1}^{L} x_i + \epsilon,$$

where $\epsilon$ is a random effect due to environment (assumed to have a mean of zero and variance $\sigma_\epsilon^2$) and

$$x_i = \begin{cases} +a & \text{if} \quad DD \\ 0 & \text{if} \quad Dd \\ -a & \text{if} \quad dd. \end{cases}$$

If $P_i$, $Q_i$ and $R_i$ are the population frequencies of individuals with genotypes $DD$, $Dd$ and $dd$ at locus $i$, and the loci are not linked so they segregate independently, the average value of the trait in the population is

$$\overline{m} = \sum_{i=1}^{L} (aP_i - aR_i) = a \sum_{i=1}^{L} (P_i - R_i),$$

and the genetic variance in the population is

$$\sigma_G^2 = a^2 \sum_{i=1}^{L} [P_i(1 - P_i) + R_i(1 + 2P_i - R_i)].$$

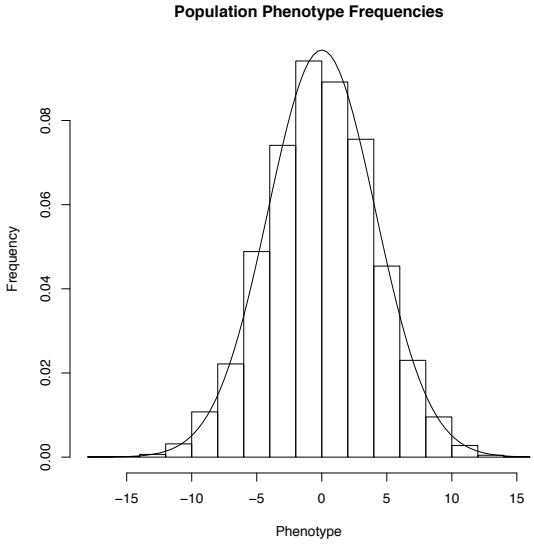


Figure 1.1 Frequency distribution of simulated phenotypic values (histogram). Probability plot for a normal distribution with mean 0 and variance 17 is the overlaid smooth line.

The total population variance of the trait is the sum of variances due to genetic factors and environments,

$$\sigma_T^2 = \sigma_G^2 + \sigma_\epsilon^2.$$

The heritability of the trait is the proportion of total phenotypic variance attributed to genetic variation,

$$H = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\epsilon^2}.$$

For example, suppose that $L = 2$, $a = 2$, $\sigma_\epsilon^2 = 1$, $P_1 = 0.2, R_1 = 0.3$, and $P_2 = 0.6, R_2 = 0.05$, then

$$\overline{m} = 2 \times [(0.2 - 0.3) + (0.6 - 0.05)] = 0.9,$$

and

$$\begin{aligned}\sigma_G &= 2^2[0.2(1 - 0.2) + 0.3(1 + 2(0.2) - 0.3)] \\ &\quad + 2^2[0.6(1 - 0.6) + 0.05(1 + 2(0.6) - 0.05)] \\ &= 3.35.\end{aligned}$$

Thus,

$$\sigma_T^2 = 3.35 + 1 = 4.45,$$

and

$$H = \frac{3.35}{3.35 + 1} = 0.77.$$

As the number of loci influencing a trait increases, the distribution of the trait in a population approaches a normal distribution with a mean and variance as given above. The frequency distribution for a trait in a population of $N = 10,000$ individuals, generated by computer simulation under the model described above, is shown in Figure 1.1. In this case, $a = 2$, and $P_i = R_i = 0.2$ for all $i = 1, 2, \ldots, 10$. The variance predicted by the above equations is $\sigma_T^2 = 17$ and the predicted mean is $\overline{m} = 0$. The mean and variance calculated for the simulated data are -0.0711 and 17.08, respectively.

## 1.3 Trait correlations between relatives

The simple model described above can be used to predict the expected degree of similarity between relatives of different degree. Observations on trait values among relatives (such as were collected by Galton, for example) can then potentially be used to predict the degree of heritability of a trait. To do so, we need a statistic for summarizing the similarities of traits between individuals with a given familial relationship. A useful measure of the association between a pair of variables $(x_i, y_i)$ is the correlation coefficient, defined as

$$\rho = \frac{\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y}\right]\left(\frac{n}{n-1}\right)}{\sqrt{\sigma_x^2 \sigma_y^2}},$$

where,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

and,

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \ \sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

are the means and variances of the variables $x$ and $y$, respectively. As an example, suppose that we measure height in 5 pairs of fathers and

| Index | Height (Father, $x$) | Height (Son, $y$) |
|---|---|---|
| 1 | 72 | 69 |
| 2 | 65 | 67 |
| 3 | 69 | 70 |
| 4 | 58 | 63 |
| 5 | 77 | 65 |

Table 1.1 *Comparative height (in inches) for a sample of fathers and sons.*

sons. The measurements (in inches) are shown in Table 1.1. The means of fathers and sons are,

$$\overline{x} = \frac{1}{5}(72 + 65 + 69 + 58 + 77) = 68.2,$$
$$\overline{y} = \frac{1}{5}(69 + 67 + 70 + 63 + 65) = 66.8,$$

the variances are,

$$\begin{aligned}\sigma_x^2 &= \frac{1}{4}[((72 - 68.2)^2 + (65 - 68.2)^2 + \\ &\quad (69 - 68.2)^2 + (58 - 68.2)^2 + (77 - 68.2)^2] = 51.7 \\ \sigma_y^2 &= \frac{1}{4}[(69 - 66.8)^2 + (67 - 66.8)^2 + \\ &\quad (70 - 66.8)^2 + (63 - 66.8)^2 + (65 - 66.8)^2] = 8.2,\end{aligned}$$

and the covariance is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i y_i &= (1/5)[72(69) + 65(67) + \\ &\quad 69(70) + 58(63) + 77(65)] \\ &= 4562.4. \end{aligned} \tag{1.1}$$

The correlation coefficient is therefore,

$$\rho = \frac{(4562.4 - 68.2 \times 66.8)(5/4)}{\sqrt{51.7 \times 8.2}} = 0.403$$

Fisher showed the expected correlation between fathers and sons to be

$$\rho_{FS} = \frac{1}{2}\left(\frac{\sigma_G}{\sigma_G + \sigma_\epsilon}\right) = \frac{1}{2}H.$$

Therefore, a rough prediction of the heritability of a trait is

$$H = 2\rho_{FS}. \tag{1.2}$$

For the example given above this would be

$$H = 2(0.403) = 0.806.$$

A sample of 5 sons and fathers is far too few to get reliable estimates of $H$ and this example is for illustration purposes only.

## 1.4 Galton's mistake: regression to the mean

In an influential paper published in the scientific journal *N*ature in 1886, Galton compared the average height of the children in a family against the average (midpoint) height of their parents. His expectation was that if height is highly heritable then taller than average parents should produce taller than average children, and so on. In making these comparisons, Galton notice an interesting phenomenon that he called "regression toward mediocrity" [now referred to as regression toward the mean (population average)]. If the difference between average heights of children versus parents is analyzed one finds that parents with more extreme heights (either short or tall) tend to have children that are more discordant from the heights of their parents and closer in height to the population mean (e.g., taller if parents are short or shorter if parents are tall). Galton was very excited by this result and developed a complex genetic explanation based on the idea that a child's height is determined by a mixture of it's parents heights and the heights of many past ancestors whose average height tends to be more similar to the population mean height. We now know that this explanation is incorrect and that regression toward the mean is a purely statistical phenomenon that has nothing to do with genetics. To illustrate regression toward the mean consider a simple simulation. Suppose that two sets of 100 random variables are simulated from a normal distribution with a mean of 10 and a standard deviation of 5. We assume that height is not inherited and therefore can treat the first 100 variables as representing the average heights of parents and the second 100 variables as representing the average heights of children; both sets of heights will be independent normally distributed random variables. Panel 1 of Figure 1.2 shows the relationship between the height of each pair of parents $P$ and the average height of their children $C$. The plot of the two variables creates a random cloud of points with no apparent association between heights. If we instead plot the difference $\Delta = P - C$ between the average height of the parents and the average height of the children againt $P$ (panel 2

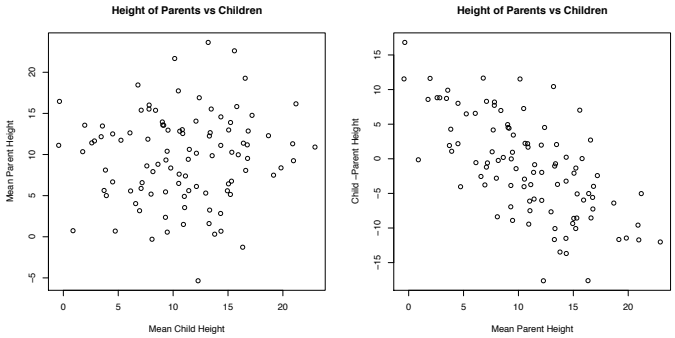


Figure 1.2 Frequency distribution of simulated phenotypic values (histogram). Probability plot for a normal distribution with mean 0 and variance 17 is the overlaid smooth line.

of Figure 1.2), a clear pattern emerges showing a greater deviation of heights for children born to either very tall or short parents – this is the phenomenon of regression to the mean observed by Galton. The simple explanation is that if we choose an extreme value of a random variable and compare it with a random value that is not chosen based on its value the deviation between the two values will be greater on average than the deviation one would expect to see between two variables neither of which is chosen to be extreme.

# 2

# The First Human Genetic Markers: ABO Blood Groups

## 2.1 Introduction

The ABO blood groups were first discovered by Karl Landsteiner in 1900 (he recieved the 1930 Nobel Prize for this work). Landsteiner identified 3 serotypes A, B and O. A fourth serotype AB was subsequently identified by Decastrello and Sturli in 1902. The serology assay mixes either *A* or *B* blood group antigens with a blood sample of unknown type and observes whether an immune cross-reaction, indicated by agglutination (clumping of cells), occurs. A cross-reaction occurs because an antigen (A or B) is present that is recognized as foreign because an individual that is a source of one of the blood samples has a blood type that does not include that antigen. If agglutination occurs with addition of both A and B antigens the sample type is O. If agglutination occurs with neither A nor B it is AB, while if it occurs only with A or only with B it is B, or A, respectively.

The blood group assay determines the phenotype of an individual but not the genotype (particular combination of alleles that determine

| Allele | $I^A$ | $I^B$ | $i$ |
|---|---|---|---|
| $I^A$ | $I^A I^A$ (A) | $I^A I^B$ (AB) | $I^A i$ (A) |
| $I^B$ | $I^A I^B$ (AB) | $I^B I^B$ (B) | $I^B i$ (B) |
| $i$ | $I^A i$ (A) | $I^B i$ (B) | $ii$ (O) |

Table 2.1 *Possible combinations of $I^A$, $I^B$ and i alleles of the ABO blood group and resultant blood group phenotype (in parentheses).*

| Allele | $I^A$ | $I^B$ | $i$ |
|---|---|---|---|
| $I^A$ | $f_{AA} = p_A^2$ | $f_{AB} = 2p_Ap_B$ | $f_{Ai} = 2p_A(1 - p_A - p_B)$ |
| $I^B$ | | $f_{BB} = p_B^2$ | $f_{Bi} = 2p_B(1 - p_A - p_B)$ |
| $i$ | | | $f_{ii} = (1 - p_A - p_B)^2$ |

Table 2.2 *Expected population frequencies of ABO genotypes as a function of population allele frequencies under Hardy-Weinberg equilibrium.*

blood type). In 1924 Felix Bernstein mathematically derived the relationship between alleles and genotypes at the ABO locus. These relationships are shown in Table 2.1. The $I^A$ and $I^B$ alleles are dominant to $i$ but are co-dominant to one another. Thus, an individual with blood group $A$ could have either of the genotypes $I^Ai$ or $I^AI^A$, whereas an individual with blood type O must have genotype $ii$, an individual with blood group AB must have genotype $I^AI^B$ and so on. If the population frequencies of the ABO blood group alleles are known, it is possible to predict the expected genotype frequencies. This relationship between allele and genotype frequencies under random mating is referred to as Hardy-Weinberg equilibrium (HWE). Most genes are in HWE in human populations. If we define $p_A$, $p_B$ and $p_i = 1 - p_A - p_B$ to be the relative population frequencies of alleles $I^A$, $I^B$ and $i$, respectively, the expected genotype frequencies under HWE are as given in Table 2.2. To obtain the expected frequency of phenotypes involving combinations of recessive and dominant alleles we must sum the genotype frequencies. For example, the expected frequency, $f_A$ of blood group phenotype A (which can result from either $I^AI^A$ or $I^Ai$) we take the sum,

$$\begin{aligned} f_A &= f_{AA} + f_{Ai} \\ &= p_A^2 + 2p_A(1 - p_A - p_B) \\ &= 2p_A(1 - p_A/2 - p_B). \end{aligned}$$

It is clear from table 2.2 that the population frequency of allele $i$ can be inferred directly from the frequency of the type O blood group as

$$p_i \approx \sqrt{f_i},$$

because the relationship between $f_i$ and $f_{ii}$ is unambiguous. However, it is not obvious how to infer population allele frequencies for $A$ and $B$ given phenotype frequencies $f_A$ and $f_B$. A very general solution to

this problem (any number of alleles, with any pattern of co-dominance) was solved by C.A.B Smith in 1955 using a procedure now referred to as the expectation-maximization (EM) algorithm. Similar methods have been used to infer haplotype phase from genotype data (a problem we consider later in the course).

Without going into statistical details, we describe here how the EM algorithm can be applied to infer population allele frequencies ($p_A$, $p_B$ and $p_i$) using population blood group phenotypes from a random sample of individuals (A, B, AB and O). This is a truly magical algorithm. Let $n_{AB}$, $n_{ii}$, $n_A$ and $n_B$ be the observed sample counts of phenotypes AB, O, A and B, respectively. We choose arbitrary initial values for the unobserved counts $\hat{n}_{AA}$, $\hat{n}_{Ai}$, $\hat{n}_{BB}$, and $\hat{n}_{Bi}$ of genotypes $I^A I^A$, $I^A i$, $I^B I^B$ and $I^B i$, respectively. The "maximization" step of the algorithm estimates the allele frequencies using the current counts by the so-called "maximum-likelihood" method. The maximum likelihood estimators of allele frequency are

$$\begin{aligned}
\hat{p}_A &= \frac{2\hat{n}_{AA} + \hat{n}_{Ai} + n_{AB}}{2n}, \\
\hat{p}_B &= \frac{2\hat{n}_{BB} + \hat{n}_{Bi} + n_{AB}}{2n}, \\
\hat{p}_i &= \frac{2n_{ii} + \hat{n}_{Ai} + \hat{n}_{Bi}}{2n},
\end{aligned}$$

where $n = n_{AB} + n_{ii} + n_A + n_B$ is the total number of individuals sampled. The "expectation" step uses the current estimates of allele frequencies to predict the expected counts for the unobserved genotypes (assuming HWE),

$$\begin{aligned}
\hat{n}_{AA} &= n_A \left( \frac{\hat{p}_A^2}{\hat{p}_A^2 + 2\hat{p}_A\hat{p}_i} \right), \\
\hat{n}_{Ai} &= n_A \left( \frac{2\hat{p}_A\hat{p}_i}{\hat{p}_A^2 + 2\hat{p}_A\hat{p}_i} \right), \\
\hat{n}_{BB} &= n_B \left( \frac{\hat{p}_B^2}{\hat{p}_B^2 + 2\hat{p}_B\hat{p}_i} \right), \\
\hat{n}_{Bi} &= n_B \left( \frac{2\hat{p}_B\hat{p}_i}{\hat{p}_B^2 + 2\hat{p}_B\hat{p}_i} \right),
\end{aligned}$$

We then return to the maximization step and predict new estimates of allele frequencies using our updated predictions of genotype counts,

| | Group O | Group A | Group B | Group AB |
|---|---|---|---|---|
| Observed | 4,578 | 4,219 | 890 | 313 |
| Predicted | 4586.8 | 4209.8 | 879.7 | 323.7 |

Table 2.3 *Observed and predicted blood group frequencies in a sample of 10,000 Londoners. Predictions are based on allele frequencies inferred using the EM algorithm and assuming Hardy-Weinberg equilibrium.*

then predict new genotype counts given our updated allele frequency estimates, and so on, until the allele frequency estimates stabilize.

To illustrate the EM algorithm we apply it to a dataset from a 1953 paper by Aird et al. that examines blood group frequencies in a sample of 10,000 people from London. The data are shown in the first row of Table 2.3. As our initial values for the unobserved counts we take

$$\begin{aligned}\hat{n}_{AA} &= 100,\\ \hat{n}_{BB} &= 100,\\ \hat{n}_{Ai} &= n_A - 100 = 4219 - 100 = 4119,\\ \hat{n}_{Bi} &= n_B - 100 = 890 - 100 = 790.\end{aligned}$$

Note that although these starting values are arbitrarily chosen, they must satisfy the constraints $n_A = n_{AA} + n_{Ai}$ and $n_B = n_{BB} + n_{Bi}$ otherwise our predicted counts will not fit the phenotype counts for the observed data ($n_A$ and $n_B$). Next, we apply the maximization step to predict the allele frequencies,

$$\begin{aligned}\hat{p}_A &= \frac{2 \times 100 + 4119 + 313}{2 \times 10000} = 0.2316\\ \hat{p}_B &= \frac{2 \times 100 + 790 + 313}{2 \times 10000} = 0.06515\\ \hat{p}_i &= \frac{2 \times 4578 + 4119 + 790}{2 \times 10000} = 0.70325\end{aligned}$$

Next we generate new predictions for the unobserved genotype counts using our current allele frequency estimates,

$$\begin{aligned}\hat{n}_{AA} &= 4219 \left( \frac{(0.2316)^2}{(0.2316)^2 + 2 \times 0.2316 \times 0.70325} \right) = 596.4962,\\ \hat{n}_{Ai} &= 4219 \left( \frac{2 \times 0.2316 \times 0.70325}{(0.2316)^2 + 2 \times 0.2316 \times 0.70325} \right) = 3622.504,\end{aligned}$$

| It | $p_A$ | $p_B$ | $p_i$ | $n_{AA}$ | $n_{Ai}$ | $n_{BB}$ | $n_{Bi}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.2316000 | 0.06515000 | 0.7032500 | 596.4962 | 3622.504 | 39.40033 | 850.5997 |
| 2 | 0.2564248 | 0.06212002 | 0.6814552 | 668.0867 | 3550.913 | 38.79694 | 851.2031 |
| 3 | 0.2600043 | 0.06208985 | 0.6779058 | 678.8881 | 3540.112 | 38.97306 | 851.0269 |
| 4 | 0.2605444 | 0.06209865 | 0.6773569 | 680.5332 | 3538.467 | 39.00855 | 850.9915 |
| 5 | 0.2606267 | 0.06210043 | 0.6772729 | 680.7842 | 3538.216 | 39.01424 | 850.9858 |
| 6 | 0.2606392 | 0.06210071 | 0.6772601 | 680.8225 | 3538.178 | 39.01512 | 850.9849 |
| 7 | 0.2606411 | 0.06210076 | 0.6772581 | 680.8283 | 3538.172 | 39.01525 | 850.9847 |
| 8 | 0.2606414 | 0.06210076 | 0.6772578 | 680.8292 | 3538.171 | 39.01528 | 850.9847 |
| 9 | 0.2606415 | 0.06210076 | 0.6772578 | 680.8294 | 3538.171 | 39.01528 | 850.9847 |
| 10 | 0.2606415 | 0.06210076 | 0.6772578 | 680.8294 | 3538.171 | 39.01528 | 850.9847 |

Table 2.4 *Allele frequency estimates and predicted genotype counts at each of 10 iterations of an EM algorithm.*

$$\hat{n}_{BB} = 890\left(\frac{(0.06515)^2}{(0.06515)^2 + 2\times 0.06515\times 0.70325}\right) = 39.40033,$$
$$\hat{n}_{Bi} = 890\left(\frac{2\times 0.06515\times 0.70325}{(0.06515)^2 + 2\times 0.06515\times 0.70325}\right) = 850.5997,$$

We now return to the maximization step and again estimate allele frequencies using the updated predictions for the genotype frequencies. The results for 10 iterations of this algorithm are shown in Table 2.4. Using the inferred population allele frequencies and assuming HWE we can predict the expected genotype counts for our sample. These are shown in the second row of Table 2.3. The predicted and expected counts are in fairly close agreement atesting to the accuracy of our inferred allele frequencies. In this case, small departures may be due to a violation of HWE for the sample.

## 2.2 Case-control studies of disease-marker association

Following the work of Bernstein in 1925 revealing that a single tri-allelic locus determined the ABO blood group types, many studies were carried out examining blood group frequencies among human populations. In a prescient paper, Lionel Penrose (1939) layed out a general framework for using association to test linkage between phenotypic traits in humans. He also identified sources of spurious asociation (not due to linkage) such as geographical substructure. In the 1950s some

TABLE III.—*Percentage Group Frequencies*

| Group | Peptic Ulcer (3,011 Cases) | | | Cancer of Stomach (2,745 Cases) | | | Cancer of Colon and Rectum (2,599 Cases) | | Cancer of Bronchus (998 Cases) | | Cancer of Breast (1,017 Cases) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Control | Disease | % Inc. or Dec. on Control | Control | Disease | % Inc. or Dec. on Control | Control | Disease | Control | Disease | Control | Disease |
| O | 47·00 | 55·40 | +17·9 | 46·78 | 42·95 | −8·2 | 46·07 | 44·79 | 46·26 | 45·49 | 46·18 | 46·12 |
| A | 40·99 | 34·67 | −15·4 | 41·38 | 46·19 | +11·6 | 41·78 | 43·63 | 41·70 | 41·28 | 41·52 | 41·00 |
| B | 8·98 | 7·44 | −17·1 | 8·79 | 7·76 | −11·7 | 8·94 | 8·66 | 8·79 | 10·72 | 8·93 | 10·03 |
| AB | 3·03 | 2·49 | −17·9 | 3·05 | 3·10 | +1·6 | 3·21 | 2·92 | 3·25 | 2·51 | 3·36 | 2·85 |

Figure 2.1 caption.

TABLE IV.—*Percentage Gene Frequencies*

| Group | Peptic Ulcer | | Cancer of Colon and Rectum | | Cancer of Bronchus | | Cancer of Breast | | Cancer of Stomach | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Control | Disease | Control | Disease | Control | Disease | Control | Disease | Control | Disease |
| O | 68·5 | 74·6 | 67·8 | 66·8 | 68·1 | 67·0 | 68·0 | 67·7 | 68·4 | 65·5 |
| A | 25·2 | 20·5 | 25·9 | 27·1 | 25·8 | 25·5 | 25·7 | 25·3 | 25·5 | 28·9 |
| B | 6·3 | 4·9 | 6·3 | 6·2 | 6·1 | 7·5 | 6·3 | 7·0 | 6·1 | 5·7 |

Figure 2.2 caption.

of the first case-control genetic association studies were carried out using serology to examine the potential association between blood group type and human disease (Aird et al 1953).

We consider here the study of Aird et al that examined associations between the ABO blood group type and 5 diseases including peptic ulcer, colon and rectal cancer, breast cancer, stomach cancer, and bronchial cancer. The basic design of an association analysis compares the frequency of one or more genetic markers in a sample of cases versus that in a sample of matched controls. If the marker frequency differs significantly between cases and controls there is an association. This may indicate that the marker (or a gene closely linked to the marker) plays a role in the development of the disease.

One can begin by examining the frequencies (of blood groups or inferred allele frequencies) in disease cases versus controls for apparent differences, which sometimes can appear quite striking. Figures 2.1 and 2.2 present the relative blood group and alleles frequencies, respectively among cases and controls for the 5 diseases mentioned above. In most cases, the frequencies are rather similar between cases and controls. However, the peptic ulcer data stands out with an apparent excess of blood group O among individuals with peptic ulcer. Note that the gene frequencies of Figure 2.2 are estimated using an algorithm similar to the EM algorithm described previously.

### 2.2.1 Relative risk and the odds ratio

The odds ratio provides a powerful method for objectively quantifying the magnitude of disease risk conferred by a factor (such as genotype) and is one of the fundamental measures used by epidemiologists. The odds ratio has been independently discovered several times during the last century, attesting to its importance and generality. Fisher (1935) studied an odds ratio statistic in the context of proportions of criminality in monozygtic versus dizygotic twins (eugenicists were interested in the heritability of criminality), Berkson (1953) derived an odds ratio (logit) in the context of logistic regression methods for analyzing dose response curves, and Woolf (1955) proposed an odds ratio as a measure to quantify the disease risk conferred by blood group type, removing the effect of blood type population frequencies. Woolf's seminal study appears to be the first application of an odds ratio in human genetic association analysis.

The odds ratio (OR) of disease is defined as

$$OR = \frac{P_1}{1-P_1} \Big/ \frac{P_2}{1-P_2} = \frac{P_1(1-P_2)}{P_2(1-P_1)}, \tag{2.1}$$

where $P_1$ and $P_2$ are the proportions of individuals exposed to the risk factor among cases and controls, respectively, and $1-P_1$ and $1-P_2$ are the proportions not exposed to the risk factor. Often the natural logarithm of the odds (the log-odds) is used instead because a change in the labels of the risk factors only changes the sign of the log-odds ($LOD$) but changes the value of the $OR$. For example, let $OR = 2$ so that $LOD = 0.693$. If we relabel the risk factors symmetrically, then $OR = 1/2$ but $LOD = -0.693$. The log-odds ($LOD$), is defined as

$$LOD = \log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_2}{1-P_2}\right).$$

A major advantage of using the odds ratio (or log-odds), rather than comparing disease incidence directly between risk-exposed and -unexposed groups, is that the odds ratio is independent of the population frequency of the risk factor. This allows one to directly compare odd ratios among populations and separate studies with potentially different frequencies for the risk factor. To see this, let $p$ be the population frequency of the risk factor and let $z_1$ and $z_2$ be the probabilities that an individual is either a case, or control, respectively, given that they are exposed to

the risk factor. The expected population proportions are,

$$\begin{aligned} P_1 &= z_1 p, \\ 1 - P_1 &= z_2(1-p), \\ P_2 &= (1-z_1)p, \\ 1 - P_2 &= (1-z_2)(1-p). \end{aligned}$$

Substituting these values into equation 2.1 above, we obtain

$$\begin{aligned} OR &= \left[\frac{z_1 p}{z_2(1-p)}\right] \Big/ \left[\frac{(1-z_1)p}{(1-z_2)(1-p)}\right], \\ &= \left(\frac{z_1}{z_2}\right) \Big/ \left(\frac{1-z_1}{1-z_2}\right), \\ &= \left(\frac{z_1}{1-z_1}\right) \Big/ \left(\frac{z_2}{1-z_2}\right), \end{aligned}$$

which is simply the ratio of the odds of being a case given an exposure to the risk factor, $z_1/(1-z_1)$, versus the odds of being a case given no exposure, $z_2/(1-z_2)$. This effectively quantifies the increased risk due to exposure to the risk factor. Clearly, if the factor does not influence risk this ratio will be 1, otherwise it will be greater than 1.

The relative risk ($RR$) is defined as the probability that an individual exposed to the risk factor develops the disease (e.g., becomes a case) divided by the probability that an unexposed individual develops the disease,

$$RR = \frac{\Pr(\text{case}|\text{exposed})}{\Pr(\text{case}|\text{unexposed})} = \frac{z_1}{z_2}.$$

The relationship between the $OR$ and $RR$ is

$$OR = RR \times \left(\frac{1-z_2}{1-z_1}\right).$$

Therefore, $OR \approx RR$ in the case that a disease is rare, so that risks for both exposed and unexposed individuals are small (e.g., $z_i << 1$, $i = 1, 2$). It is not always possible to estimate the $RR$ directly in case-control studies. The relevant ratio of population parameters is

$$\begin{aligned} RR &= \frac{\Pr(\text{exposed}|\text{case})}{\Pr(\text{unexposed}|\text{case})} \times \frac{\Pr(\text{unexposed})}{\Pr(\text{exposed})} \\ &= \left(\frac{p_C}{1-p_C}\right) \times \left(\frac{1-p}{p}\right), \end{aligned}$$

where $p_C$ is the frequency of the risk factor among cases and $p$ is the

overall population frequency of the risk factor (usually unknown). In much of the human genetics literature the terms relative risk and odds ratio are used interchangeably to refer to what here we call the odds ratio.

### 2.2.2 Odds ratio estimators

A straightforward estimator of the *OR* (Woolf, 1955) uses the observed sample proportions to estimate the population proportions, applying equation 2.1 above. Table 2.5 presents a contingency table of the out-

| | Risk Factor | | |
|---|---|---|---|
| Disease | + | - | Total |
| + | $a$ | $b$ | $n_+$ |
| - | $c$ | $d$ | $n_-$ |

Table 2.5 *Contingency table of possible outcomes for a case-control study of a binary risk factor. Plus and minus signs indicate the presence or absence of either the risk factor (row) or the disease (column).*

comes for a case-control association study of a binary disease trait. The estimator of Woolf (1955) is

$$\hat{OR} = \frac{a \times d}{b \times c}. \tag{2.2}$$

One approach for applying these formulae to biallelic SNPs is to partition genotypes into binary classes, for example if we label the alleles 1 and 2 we could consider 11 versus 12 or 22, and so on (Thomson, 1981). This quantity is referred to as the genotype relative risk. Here, we will compare risks among phenotypes (e.g., bloog group A versus O, etc) because the genotypes are unknown due to the dominance of A and B over O. To illustrate, we apply equation 2.2 to the analysis of blood group data for two diseases included in the case-control study of Aird el al. (1954), colon cancer and peptic ulcer. Tables 2.6 and 2.7 show the counts for each disease in a study of individuals from London hospitals (combining counts for males and females). The odds ratio for colon

| | Blood Group | | |
|---|---|---|---|
| | O | A | Total |
| Case | 665 | 676 | 1341 |
| Control | 4578 | 4219 | 8797 |

Table 2.6 *Combined male and female counts of A and B blood types among patients with colon cancer (case) and normal individuals (control) from London hospitals.*

cancer is

$$\hat{OR} = \frac{665 \times 4219}{676 \times 4578} = 1.10,$$

and the odds ratio for peptic ulcer is

$$\hat{OR} = \frac{911 \times 4219}{579 \times 4578} = 1.45,$$

| | Blood Group | | |
|---|---|---|---|
| | O | A | Total |
| Case | 911 | 579 | 1490 |
| Control | 4578 | 4219 | 8797 |

Table 2.7 *Combined male and female counts of A and B blood types among patients with peptic ulcer (case) and normal individuals (control) from London hospitals.*

Ultimately, a statistical hypothesis test is needed to decide whether observed frequency differences are significant, given the sample sizes, etc, used in the study. With large sample sizes, a very subtle difference in frequency may be significant whereas with small samples even quite dramatic differences my be within the realm of deviations expected due to sampling effects alone. We consider two hypotheses: the null hypothesis (which we aim to test) and the alternative hypothesis (which can be rather vague). In a simple association analysis, the null hypothesis is that the $OR$ is 1, while under the alternative hypothesis $OR \neq 1$. We can test this hypothesis by constructing a 95% confidence interval for

the estimate of $OR$ and checking whether the interval includes 1, if so we accept the null hypothesis that $OR = 1$, otherwise we reject the null hypothesis at the $\alpha = 1 - 0.95 = 0.05$ level of significance. Woolf (1955) developed an approximate method for inferring the standard deviation of $OR$,

$$\hat{\sigma} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

We can use this equation to infer an approximate 95% confidence interval for OR,

$$\hat{OR} \pm 1.96 \times \hat{OR} \times \hat{\sigma},$$

For the example considering relative frequency of A and O blood groups among colon cancer patients versus controls the standard deviation is

$$\hat{\sigma} = \sqrt{\frac{1}{556} + \frac{1}{676} + \frac{1}{4578} + \frac{1}{4219}} = 0.059,$$

and the 95% confidence interval is

$$\hat{OR} \in (1.10 - 1.96 \times 0.059 \times 1.10, 1.10 + 1.96 \times 0.059 \times 1.10) = (0.97, 1.23).$$

Thus the CI for the odds ratio of blood types among cases and controls includes 1 and we fail to reject the null hypothesis of no increased cancer risk among O versus A blood groups. Considering relative frequency of A and O blood groups among peptic ulcer patients versus controls the standard deviation is

$$\hat{\sigma} = \sqrt{\frac{1}{911} + \frac{1}{579} + \frac{1}{4578} + \frac{1}{4219}} = 0.057,$$

and the 95% confidence interval is

$$\hat{OR} \in (1.45 - 1.96 \times 0.057 \times 1.45, 1.45 + 1.96 \times 0.057 \times 1.45) = (1.28, 1.61).$$

Thus the CI for the odds ratio of blood groups among cases and controls excludes 1 and we reject the null hypothesis of no increased risk of peptic ulcer among O versus A blood groups. Because $OR > 1$ there is significant evidence for an increased risk of peptic ulcer among individuals with the O blood group type.

# 3
# Human Genomic Variation

Humans are diploid, meaning that a normal individual possesses two copies (homologues) of each of 22 autosomal chromosomes. A pair of homologous chromosomes will contain essentially all the same genes, regulatory regions, etc, arranged in the same order on the chromosome, but the DNA sequences on each chromosome vary in subtle ways due to insertions, deletions, point mutations, etc that have accumulated over a long period of human evolutionary history. The position on a chromosome where a particular genetic variant occurs is referred to as a **locus** and the set of variants at that position (locus) found in a population are referred to as **alleles**. For historical reasons, genetic variants that can be analyzed using DNA sequencing or other techniques are referred to as **genetic markers**.

## 3.1 Single Nucleotide Polymorphisms (SNPs)

The most common genetic variants in the human genome are single nucleotide changes arising by point mutation, commonly referred to as single nucleotide polymorphisms (SNPs). A sample of 3 aligned sequences from homologous chromosomes is shown below with one nucleotide that varies at the sixth nucleotide position,

```
5' AATTCGCCT 3'
   AATTCCCCT
   AATTCCCCT
```

In this example, `C` is replaced by `G` in the first sequence. SNPs are typically identified by 5′ and 3′ flanking sequences that can be used to develop a genotyping method specific for the SNP. The physical location

of the SNP in the human genome is also usually known. Because the mutation rate of nuclear DNA is very low (on the order of $10^{-9}$) most SNPs are a result of a single point mutation and therefore only two alleles exist in the population. In the above example, the alleles are `G` and `C`. The allele that is least frequent in the population is normally referred to as the **minor allele**. A site in a DNA sequence is **polymorphic** if the minor allele frequency is greater than some predefined frequency. Often a frequency of 0.05 is used to choose sites for inclusion in SNP databases.

## 3.2 Genotypes, Haplotypes and Diplotypes

The combination of alleles found in an individual at a particular genetic marker locus comprise the individual **genotype**. For example, if at a specific SNP site an individual possessed a `T` on one homologous chromosome and a `C` on the other, the genotype of that individual would be `C/T`. Modern genotyping methods can be used to type thousands, or millions, of SNPs for a single individual. The **multi-locus genotype** is the combination of alleles observed at 2 or more SNPs in a single individual. A **haplotype** is a distinct combination of alleles at 2 or more genetic marker loci that are found on a particular chromosome. A **diplotype** is the pair of haplotypes residing on the homologous chromosomes of an individual. For example, a diplotype of 3 SNPs is

```
5'...A.........T.............G..3'
5'...T.........T.............C..3'
```

where a . indicates a non-polymorphic intervening site in the sequence. Each line above represents a particular haplotype and the multi-locus genotype is `A/T,T/T,G/C`.

## 3.3 Summarizing Human Genomic Variation

### 3.3.1 Allele frequency

An important statistic summarizing the variation at a genetic marker is the allele frequency. Suppose that a population is comprised of $N$ diploid individuals. Let $X_i$ be the genotype of individual $i$ at a particu-

lar genetic locus. The frequency of allele $A$ at the locus is defined as

$$p_A = \frac{1}{2N}\sum_{i=1}^{N} I_A(X_i),$$

where

$$I_A(X_i) = \begin{cases} 1 & \text{if} \quad X_i = \texttt{A/}\cdot \\ 2 & \text{if} \quad X_i = \texttt{A/A} \\ 0 & \text{otherwise} \end{cases}$$

In most cases, we do not sample the entire population and we therefore estimate the population allele frequency using a random sample of $n$ individuals,

$$\hat{p}_A = \frac{1}{2n}\sum_{i=1}^{n} I_A(X_i),$$

where the "hat" symbol is used to indicate that this is an estimate of the population allele frequency "parameter." To illustrate the calculation of allele frequency we analyze data on human SNP polymorphisms collected by the human HapMap project and available online from the dbSNP database. We focus on RefSNP rs999991 which is located on chromosome 14. The European HapMap sample for this marker comprises 60 unrelated individuals (120 chromosomes). The major allele is `C` and the minor allele is `T`. The genotype counts are `C/T` : 4 and `T/T` : 56. The estimates of the allele frequencies of `C` and `T`, respectively, are

$$p_C = \frac{(56 \times 2) + (4 \times 1)}{2 \times 60} = 0.967,$$

and

$$p_T = \frac{4 \times 1}{2 \times 60} = 0.033 = 1 - p_C.$$

### 3.3.2 Genotype Frequencies and Hardy-Weinberg Proportions

A fundamental result in population genetics is that the expected genotype proportions at a locus in a large randomly-mating population rapidly converge to those give by a simple function of the gene frequencies known as Hardy-Weinberg equilibrium (HWE) after its co-discoverers. The expected proportions under HWE are:

| `A/A` | `A/T` | `T/T` |
|---|---|---|
| $p_A^2$ | $2p_A(1-p_A)$ | $(1-p_A)^2$ |

This result can be understood by noting that with random mating we choose a pair of gametes at random from the population of potential gametes to form an individuals genotype. Each gamete carries allele A with probability $p_A$. The A/A genotype is therefore obtained with probability $p_A \times p_A = p_A^2$ (because the probability that two independent events both occur is equal to the product of their probabilities). In a large population, the expected frequency of A/A genotypes is proportional to the probability that a random individual possesses this genotype. The frequency of heterozygotes is the probability of either ordered genotype A/T or T/A because it does not concern us whether A, for example, is received from the mother or the father, so the probability of a heterozygote is $p_A(1-p_A) + (1-p_A)p_A = 2p_A(1-p_A)$. Using the allele frequencies estimated previously at SNP marker rs999991 for the HapMap sample of Europeans we can calculate the expected genotype proportions (and counts) under HWE for this sample. The expected proportions are

$$\begin{aligned} f(\mathtt{C/C}) &= p_C^2 = (0.967)^2 = 0.935, \\ f(\mathtt{C/T}) &= 2p_C(1-p_C) = 2 \times 0.967 \times 0.033 = 0.064, \\ f(\mathtt{T/T}) &= (1-p_C)^2 = (0.033)^2 = 0.001, \end{aligned}$$

and the expected counts are

$$\begin{aligned} \mathtt{E}(n_{CC}) &= n \times f(\mathtt{C/C}) = 60 \times 0.935 = 56.1, \\ \mathtt{E}(n_{CT}) &= n \times f(\mathtt{A/T}) = 60 \times 0.064 = 3.84, \\ \mathtt{E}(n_{TT}) &= n \times f(\mathtt{T/T}) = 60 \times 0.001 = 0.06. \end{aligned}$$

The difference between the genotype counts predicted under HWE and those observed (e.g., $n_{CC} = 56, n_{CT} = 4, n_{TT} = 0$) appears very slight. We can use a $\chi^2$ test to further examine the goodness of fit of the observed to the expected genotype proportions. The $\chi^2$ test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(Obs_i - Exp_i)^2}{Exp_i},$$

where $k$ is the number of categories (in this case there are 3 categories corresponding to the 3 possible genotypes), $Obs_i$ is the observed number in category $i$ and $Exp_i$ is the expected number in category $i$. The degrees of freedom (df) for the $\chi^2$ test is the difference between the number of free observations and the number of parameters estimated from the data under the null hypothesis (HWE). In this case, only one

parameter, the allele frequency $p_C$ must be estimated and there are two free observations (the counts for any two of the genotypes). The counts for the third genotype are determined by the other two counts and the sample size, which is fixed by the experimenter. Thus the number of degrees of freedom is $2 - 1 = 1$. The test statistic should therefore follow a $\chi^2$ distribution with 1 df. The tail probability $\alpha$ specifies the probability that a value as large as our calculated value is observed under the null hypothesis. The tail probability can be obtained by looking up the $\chi^2$ value in a book of statistical tables, by calculating it directly in a software package such as R, or by using an online calculator, for example the calculator at
`http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html`.
In this case, the null hypothesis specifies that the proportions fit those expected under HWE. Thus, we accept (fail to reject) the null hypothesis that genotype frequencies are in HWE when this probability is large (e.g., $\alpha > 0.05$). For the European HapMap data we have,

$$\chi^2 = \frac{(56 - 56.1)^2}{56.1} + \frac{(4 - 3.89)^2}{3.84} + \frac{(0 - 0.06)^2}{0.06} = 0.07.$$

The tail probability, which is the probability of observing a value of $\chi^2$ at least as large as 0.07 under the null hypothesis, is $\alpha = 0.79$ and so we accept the null hypothesis that the genotype proportions fit those expected under HWE.

## 3.4 Genotype Frequencies with Inbreeding

The previous results for Hardy-Weinberg genotype proportions are based on an assumption that the choice of a mate does not depend on the genotype of the locus under investigation. If individuals tend to mate with relatives (inbreeding) or they come from separate populations that exchange few migrants (population subdivision), pairs of alleles will be more similar at the locus than would be expected under HWE. We now consider a slightly more realistic model of genotype proportions that allows for such effects by adding an **inbreeding coefficient** denoted as $F$ that represents the probability that a pair of alleles are **identical by descent**, meaning that they are descended from a recent common ancestor. Inbreeding and population subdivision generate positive values of $F$, while random mating results in $F = 0$. With inbreeding, the geno-

type proportions are

$$\begin{array}{ccc} \texttt{A/A} & \texttt{A/T} & \texttt{T/T} \\ (1-F)p_A^2 + Fp_A & (1-F)2p_A(1-p_A) & (1-F)(1-p_A)^2 + F(1-p_A) \end{array}$$

Based on these formulas, it is clear that two possible explanations for a departure from Hardy-Weinberg equilibrium are inbreeding and population subdivision (there are also other possible explanations, as we shall see later, such as overdominant selection).

## 3.5 Population Subdivision and the Wahlund Effect

The **Wahlund effect** refers to the excess proportion of homozygotes observed in a subdivided population versus a randomly mating population. We illustrate the origin of this effect using a simple case of two populations that exchange no migrants, with allele frequencies $p_1$ and $p_2$, respectively, for allele $A$ at a biallelic locus. Table 3.1 gives the expected genotype proportions in each population, the expected proportion in the "pooled" (or admixed) population (a population made up of a mixture of individuals from each population), and the expected genotype proportions under HWE (ignoring the underlying population subdivision). Using these results, it is possible to show directly that

| Genotype | A/A | A/a | a/a |
|---|---|---|---|
| Population 1 | $p_1^2$ | $2p_1(1-p_1)$ | $(1-p_1)^2$ |
| Population 2 | $p_2^2$ | $2p_2(1-p_2)$ | $(1-p_2)^2$ |
| Pooled | $(p_1^2+p_2^2)/2$ | $p_1(1-p_1)+p_2(1-p_2)$ | $[(1-p_1)^2+$ $(1-p_2)^2]/2$ |
| No subdivision | $\overline{p}^2$ | $2\overline{p}(1-\overline{p})$ | $(1-\overline{p})^2$ |

Table 3.1 *Expected genotype proportions in each or two populations (Populations 1 and 2), in a population that is an admixture of the two populations (Pooled), and the genotype proportions that would be expected in a population with the same average allele frequencies and no admixture (No subdivision). Note that $\overline{p}$ denotes the average allele frequency in the admixed population, $\overline{p} = (p_1 + p_2)/2$.*

population subdivision has an effect similar to inbreeding (e.g., increasing the proportion of homozygotes). Using the formulas given above, we represent the expected proportion of homozygotes in the admixed population by assuming that the effect will be similar to inbreeding, namely

$$f(A/A) = \overline{p}^2(1-F) + \overline{p}F,$$

and we then set this equal to the expected proportion in the admixed population, from table 3.1 above,

$$\overline{p}^2(1-F) + \overline{p} = (p_1^2 + p_2^2)/2.$$

We then solve for $F$ to determine the inbreeding coefficient in this situation,

$$F = \frac{(p_1^2 + p_2^2) - 2\overline{p}^2}{2(\overline{p} - \overline{p}^2)} = \frac{(p_1^2 + p_2^2)/2 - \overline{p}^2}{\overline{p}(1-\overline{p})} = \frac{var(p)}{\overline{p}(1-\overline{p})}, \tag{3.1}$$

where we have made use of the standard result from statistics that "the average of the square minus the square of the average" equals the variance,

$$(p_1^2 + p_2^2)/2 - \overline{p}^2 = var(p).$$

The variance is always a positive number, as is the product in the denominator, because $\overline{p} < 1$, and the inbreeding coefficient is therefore also positive leading to an excess of homozygotes in the admixed population over what would be expected under random mating. The formula for $F$ given in equation 3.1 is commonly referred to as the fixation index and can be used as a measure of the degree of population subdivision.

# 4

# Population subdivision, haplotype inference and linkage disequilibrium

## 4.1 Subdivision Between Han Chinese and Europeans

To illustrate how the distribution of genotype proportions in a population sample can be used to detect population subdivision we again consider SNP marker rs999991 from the dbSNP database. The genotype counts in a sample of $n = 43$ unrelated Han Chinese are $C/C : 14$, $C/T : 20$, and $T/T : 9$. The allele frequencies in the sample are,

$$\begin{aligned} p_C &= \frac{2(14) + 20}{2(43)} = \frac{48}{86} = 0.558, \\ p_T &= 1 - p_C = 0.442, \end{aligned}$$

and the expected genotype counts are

$$\begin{aligned} \mathtt{E}(n_{CC}) &= 43 \times (0.558)^2 = 13.389, \\ \mathtt{E}(n_{CT}) &= 43 \times 2(0.558)(0.442) = 21.211, \\ \mathtt{E}(n_{TT}) &= 43 \times (0.442)^2 = 8.401. \end{aligned}$$

The $\chi^2$ goodness of fit test statistic is

$$\chi^2 = \frac{(14 - 13.389)^2}{13.389} + \frac{(20 - 21.211)^2}{21.211} + \frac{(9 - 8.401)^2}{8.401} = 0.13973.$$

The probability of a $\chi^2$ value at least as great as this under the null hypothesis that proportions are in HWE is $\alpha = 0.709$ and we therefore fail to reject the null hypothesis. Next, we pool the Han Chinese sample with the European sample we examined previously. We now have $n = 103$ unrelated individuals with genotype counts $C/C : 56 + 14 = 70$, $C/T : 4 + 20 = 24$, and $T/T : 0 + 9 = 9$. The pooled allele frequencies

are

$$\overline{p}_C = \frac{2(70) + 24}{2(103)} = 0.7961,$$
$$\overline{p}_T = 1 - \overline{p}_C = 0.2039,$$

and the expected genotype counts are

$$\mathtt{E}(n_{CC}) = 103 \times (0.7961)^2 = 65.282,$$
$$\mathtt{E}(n_{CT}) = 103 \times 2(0.7961)(0.2039) = 33.437,$$
$$\mathtt{E}(n_{TT}) = 103 \times (0.2039)^2 = 4.281.$$

The $\chi^2$ goodness of fit test statistic is

$$\chi^2 = \frac{(70 - 65.282)^2}{65.282} + \frac{(24 - 33.437)^2}{33.437} + \frac{(9 - 4.281)^2}{4.281} = 8.206.$$

The probability of a $\chi^2$ value at least as great as this under the null hypothesis that proportions are in HWE is $\alpha = 0.004$ and we therefore reject the null hypothesis at the $\alpha = 0.004$ level of significance. The genotype proportions do not fit those expected under HWE and the most likely explanation is population subdivision. We can also calculate the fixation index for these data. The between-population variance of allele frequency is

$$var(p) = \frac{0.558^2 + 0.967^2}{2} - 0.7961^2 = 0.08364,$$

and the fixation index is

$$F = \frac{0.08364}{0.7961(1 - 0.7961)} = 0.515.$$

Thus, there is a large degree of genetic differentiation between the two populations at this locus as measured by the fixation index.

## 4.2 Marker Heterozygosity

A simple summary statistic for quantifying the variation of a genetic marker is the **heterozygosity**, which is the proportion of heterozygotes. Heterozygosity can be calculated from a sample of genotypes in several ways. The direct estimator of heterozygosity simply counts the proportion of heterozygotes present in the sample,

$$\hat{h}_1 = \frac{1}{n} \sum_{i=1}^{n} I_{het}(X_i), \tag{4.1}$$

where

$$I_{het}(X_i) = \begin{cases} 1 & \text{if} \quad X_i = \texttt{Y/Z}, \\ 0 & \text{otherwise}, \end{cases}$$

and $\texttt{Y} \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ and $\texttt{Z} \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ are any two nucleotides such that $\texttt{Y} \neq \texttt{Z}$. The allele frequency estimator of heterozygosity assumes Hardy-Weinberg proportions to estimate heterozygosity,

$$\hat{h}_2 = 1 - \sum_{i=1}^{k} p_i^2, \tag{4.2}$$

where $k$ is the number of alleles at a locus (for most SNPs $k = 2$). The allele frequency estimator will be more accurate than the direct estimator if the genotype frequencies are in HWE in the population. Alternatively, the direct estimator may be more accurate if the proportions deviate sufficiently from HWE. However, very often the two estimators give highly similar results. For the European HapMap sample, at SNP rs999991, the direct estimator gives,

$$\hat{h}_1 = \frac{4}{60} = 0.067,$$

and the indirect estimator gives,

$$\hat{h}_2 = 1 - (0.967)^2 - (0.033^2) = 0.064.$$

In this case, the two estimates vary only slightly at the third decimal place.

## 4.3 Inference of Haplotype Phase

Current genotyping technologies provide multi-locus genotypes but do not directly provide the **haplotype phase**, that is, the diplotype describing the alleles that co-occur on each homologous chromosome. In many cases, it is possible to determine the haplotype phase from the sample of multilocus genotypes. Here, we consider one simple technique that infers phase by examining the genotypes of close relatives (in our example, a "trio" of 2 parents and 1 child). Later, we will consider more sophisticated methods for phase inference using either family pedigrees or population samples of unrelated individuals.

### 4.3.1 Genotypes to diplotypes: A one to many mapping

A multilocus genotype is typically compatible with many possible haplotypes. For example, if an individual is heterozygous for two linked SNP loci there are two possible distinct combinations of diplotypes,

```
    Genotypes
-------------------
SNP 1         SNP 2
A/C           G/T
-------------------
Possible Diplotypes
A-G/C-T     A-T/C-G
```

If one of the genotypes is homozygous the diplotype (haplotype phase) is determined,

```
    Genotypes
-------------------
SNP 1         SNP 2
A/C           T/T
-------------------
Possible Diplotypes
A-T/C-T
```

Even if an individual is heterozygous for a given locus, if additional relatives are examined it is likely that one or more of the relatives are homozygous at the locus (and phase might therefore be indirectly established). The following analysis of two-locus genotypes from the members of a nuclear family illustrates this principle,

```
Mother Genotypes        Father Genotypes
----------------        ----------------
A/A,G/G                 C/A,T/T

Mother Diplotypes       Father Diplotypes
-----------------       -----------------
A-G/A-G                 C-T/A-T

Child Genotypes
---------------
A/C,G/T
```

```
Possible Child Diplotypes (ignoring parents)
--------------------------------------------
A-G/C-T    C-G/A-T

Possible Child Diplotypes (considering parents)
-----------------------------------------------
A-G/C-T
```

In the above example, the diplotype of the child (whose genotypes are heterozygous at both loci) cannot be directly inferred. The child's diplotype can be indirectly inferred, however, by considering the parental genotypes. This process is known as **phase inference**. Haplotype information is needed to test for non-random associations of alleles on chromosomes, due to linkage or other factors. The non-random assortment of alleles onto chromosomes is referred to as "linkage disequilibrium." Measures of linkage disequilibrium will be the subject of the next section.

## 4.4 Linkage Disequilibrium

In the absence of recombination, alleles located on the same chromosome are co-transmitted to offspring. Here we consider a simple statistic that is often used to quantify the non-random assortment of alleles onto chromosomes (e.g., non-random frequencies of haplotypes). If two alleles are sufficiently far apart on a chromosome, recombination occurs at each generation (meiosis) and the alleles assort independently to form genotypes in offspring (e.g., Mendel's second law applies). Alleles that assort independently are in **linkage equilibrium** and those that are non-randomly assorted onto chromosomes are in **linkage disequilibrium**. To Illustrate, consider a pair of biallelic genetic markers located on the same chromosome. Marker locus 1 has alleles $A$ and $a$, marker locus 2 has alleles $B$ and $b$. The possible haplotypes are

```
A-B
a-B
a-b
A-b
```

We denote the frequency of haplotype A-B as $p_{AB}$ and so on. The marginal (or total) frequency of allele $A$ is obtained by summing up the frequen-

cies of all haplotypes that contain allele $A$ at locus 1,

$$p_A = p_{AB} + p_{Ab},$$

and the marginal frequencies of other alleles are obtained similarly. The **disequilibrium coefficient** is then defined as

$$D = p_{AB} - p_A p_B. \tag{4.3}$$

To obtain a statistic that varies between -1 and +1, equation 4.3 above is normalized by dividing by its maximum possible value to obtain,

$$D' = D/D_{max},$$

where

$$D_{max} = \begin{cases} \min(p_A[1-p_B], [1-p_A]p_B) & \text{if} \quad D \geq 0 \\ \min(p_A p_B, [1-p_A][1-p_B]) & \text{if} \quad D < 0 \end{cases}$$

Because the sign of the statistic $D'$ depends on the labeling of alleles, which is arbitrary, the absolute value $|D'|$ is most often used. To illustrate the use of this statistic, consider a sample with $p_{AB} = 0.6$, $p_{aB} = 0.1$, $p_{ab} = 0.1$ and $p_{Ab} = 0.2$. The marginal allele frequencies are

$$\begin{aligned} p_A &= 0.6 + 0.2 = 0.8 \\ p_B &= 0.6 + 0.1 = 0.7 \\ p_a &= 0.1 + 0.1 = 0.2 \\ p_b &= 0.1 + 0.2 = 0.3, \end{aligned}$$

the disequilibrium coefficient $D$ is

$$D = 0.6 - 0.8 \times 0.7 = 0.04,$$

and the normalized disequilibrium coefficient $D'$ is

$$D' = \frac{0.04}{\min(0.24, 0.14)} = \frac{0.04}{0.14} = 0.286.$$

In this case, the observed frequency ($p_{AB} = 0.6$) of A-B is greater than expected under random assortment ($p_A \times p_B = 0.8 \times 0.7 = 0.56$) and there is positive linkage disequilibrium of the alleles. Two extreme cases of $D'$ are worth considering. First, there is the case of complete disequilibrium. An example is $p_{AB} = p_{ab} = 0.5$ and $p_{aB} = p_{Ab} = 0$,

so only two of the four possible haplotypes are observed. In this case, $p_A = p_B = p_b = p_a = 0.5$, and

$$D' = \frac{0.5 - 0.5^2}{0.5^2} = \frac{1 - 0.5}{0.5} = 1.$$

The other extreme is complete equilibrium. An example is $P_{AB} = p_{Ab} = p_{aB} = p_{ab} = 0.25$. In this case, $p_A = p_B = p_b = p_a = 0.5$, and

$$D = p_{AB} - p_A p_B = 0.25 - 0.5 \times 0.5 = 0,$$

and therefore $D' = 0$. The statistic $|D'|$ is often used to summarize linkage disequilibrium across the genome in human population samples. There is a negative relationship between $|D'|$ and physical distance along a chromosome because markers spaced at greater intervals along a chromosome experience more recombination events, on average, per generation. The patterns of LD across the human genome are somewhat variable among populations and some regions of the genome have significantly lower LD over a similar physical distance due to the presence of recombination hotspots. Average levels of LD across the genome are related to the time that has elapsed since all humans last shared a common ancestral genome.

# 5
# Human Population Genetics

## 5.1 Allele Frequency Change in Populations

Classical models of population genetic evolution focus on describing the change in allele frequencies from one generation to the next under the influence of forces such as migration, mutation, selection and genetic drift. The bulk of the theory in this area was developed from the 1930s to the 1970s. This theory has taken on new relevance with the huge increase in population genetic data during the last several decades. Here, we consider some simple population genetic results pertinent to studies of human genetic variation.

### 5.1.1 Continuous approximation of allele frequency

Define the relative population frequency of allele $A$ to be

$$p_A = \frac{n_A}{2N},$$

where $n_A$ is the number of chromosomes in the population carrying allele $A$ and $N$ is the diploid population size. The smallest possible change in allele frequency is $1/(2N)$ (e.g., an increase, or decrease, of one copy of the allele). If $N = 2$, for example, then 0.25 is the smallest possible change of allele frequency and $p_A$ takes possible values $p_A \in \{0, 0.25, 0.5, 0.75, 1\}$. If $N = 100$ then $1/(200) = 0.005$ is the smallest possible change of frequency and $p_A$ assumes a finer range of values. Finally, in the limit as $N \to \infty$ the smallest possible change in frequency approaches 0, $\lim_{N\to\infty} 1/(2N) = 0$ and $p_A$ appears continuous. This is the basis for modeling the change of allele frequency in large populations as a continuous variable. The formal "diffusion theory" for this approximation is beyond the scope of this course and we

will assume that the approximation is valid for the situations we consider without providing any formal proof.

The first set of results that we consider are for the so-called **deterministic models**. These models are valid when the population is very large so that the allele frequency changes under the effects of selection, migration, etc, in a completely predictable way (e.g., stochastic events such as random variations in number of offspring per individual, etc, can be ignored). In small populations, random effects are important and we will later consider a model known as "genetic drift" that accounts for such effects. For simplicity, all our analyses will assume that the species under consideration have discrete non-overlapping generations. This approximates the situation for human populations if we set the generation time to be approximately 20 years.

### 5.1.2 Migration between populations

Suppose that two large populations, both of size $N$, exchange migrants in a symmetrical manner such that a proportion $m$ of the individuals in each population are replaced by migrants from the other population in each generation. Consider a biallelic locus and let $p_1^{(t)}$ and $p_2^{(t)}$ be the frequencies of allele $A$ in populations 1 and 2, respectively, at generation $t$. Under this model, the average allele frequency in population 1 at the next generation is given by the equation

$$p_1^{(t+1)} = p_1^{(t)}(1-m) + p_2^{(t)}m,$$

and the average allele frequency for population 2 at the next generation is

$$p_2^{(t+1)} = p_2^{(t)}(1-m) + p_1^{(t)}m.$$

These equations can be understood by noting that in each generation a fraction $(1-m)$ of individuals are non-migrants and therefore have allele frequencies identical to those of the previous generation in that population and a fraction $m$ of individuals are migrants and therefore have allele frequencies identical to those of the alternative population at the previous generation. These equations are "iterative" meaning that the allele frequency at any point in the future can be obtained by successive applications of the equations for single generation change. For example, if the initial frequencies are $p_1^{(0)}$ and $p_2^{(0)}$ then the frequencies

after 2 generations are

$$
\begin{aligned}
p_1^{(1)} &= p_1^{(0)}(1-m) + p_2^{(0)}m, \\
p_2^{(1)} &= p_2^{(0)}(1-m) + p_1^{(0)}m, \\
p_1^{(2)} &= p_1^{(1)}(1-m) + p_2^{(1)}m, \\
p_2^{(2)} &= p_2^{(1)}(1-m) + p_1^{(1)}m.
\end{aligned}
$$

The **equilibrium allele frequency** is achieved when the allele frequencies are no longer changing from one generation to the next (this is also referred to as the stationary or steady-state frequency). We can solve for the change in allele frequency per generation in population 1 as follows

$$\Delta p_1 = p_1^{(t+1)} - p_1^{(t)} = m(p_1^{(t)} - p_2^{(t)}).$$

It is clear that the frequency is no longer changing (e.g., $\Delta p_1 = 0$) if either there is no migration ($m = 0$) or the allele frequencies are equal in the two populations ($p_1^{(t)} = p_2^{(t)}$). Thus, with ongoing migration the populations have equal allele frequencies at equilibrium. The number of generations required to reach equilibrium depends on the migration rate; with a higher migration rate equilibrium is achieved more quickly.

### 5.1.3 Mutation

The mutation process in humans is very complex. Later, we will consider models of DNA substitution that are intended to be realistic for humans. Here, we consider a highly simplified model of mutation to get some qualitative feeling for the importance of mutation in modifying gene frequencies. Suppose that two possible alleles exist at a locus $A$ and $a$. The rate of mutation per generation from allele $A \rightarrow a$ is $\mu$ and the rate of mutation from $a \rightarrow A$ is $\nu$. The frequency of $A$ at generation $t$ is $p_A^{(t)}$. The frequency of $A$ at the next generation under this model of mutation pressure is

$$p_A^{(t+1)} = p_A^{(t)}(1-\mu) + (1-p_A^{(t)})\nu.$$

This equation can be understood by noting that the current proportion of $A$ alleles is $P_A^{(t)}$ and a fraction $(1-\mu)$ of these do not experience mutation and are $A$ alleles in the next generation. Conversely, the current proportion of $a$ alleles is $(1-p_A^{(t)})$ and a fraction $\nu$ of these $a$ alleles mutate to become $A$ alleles in the next generation. To determine the

equilibrium frequency of allele $A$ we solve for the expected change in allele frequency per generation under mutation pressure,

$$\Delta p_A = -p_A^{(t)}\mu + (1 - p_A^{(t)})\nu.$$

The allele frequency is stable if $\Delta p_A = 0$, namely if,

$$p_A^{(t)}\mu = (1 - p_A^{(t)})\nu.$$

Solving the above equation for $p_A$ yields the equilibrium frequency, $p_A*$, of the $A$ allele,

$$p_A* = \frac{\nu}{\nu + \mu}.$$

Thus, at equilibrium the frequency of allele $A$ is equal to the relative rate of mutation to allele $A$ from allele $a$.

### 5.1.4 Selection

Natural selection is the major force underlying adaptive evolution. There are many good examples of current, and past, episodes of selection in humans. Here, we consider the classical single locus theory of natural selection in large populations. Despite the simplicity of the model this theory has serious applications in studying many human genetic polymorphisms. Table 5.1 provides the parameters of a selection model for a single genetic locus with two alleles $A$ and $a$. The mean fitness of the population is the proportion of the population that survive selection,

$$\overline{w} = w_{AA}p_A^2 + w_{Aa}2p_A(1 - p_A) + w_{aa}(1 - p_A)^2.$$

We now derive an iterative equation for the change in allele frequency per generation under selection pressure. The genotype proportions among adults are given in the third row of Table 5.1. Individuals that are $A/A$ produce entirely $A$ gametes, whereas individuals that are $A/a$ produce $1/2$ $A$ gametes on average. Thus, the proportion of $A$ gametes at the next generation of mating is

$$f(A) = \frac{p_A^2 w_{AA}}{\overline{w}} + \frac{1}{2}\frac{2p_A(1 - p_A)w_{Aa}}{\overline{w}} = \frac{p_A(p_A w_{AA} + (1 - p_A)w_{Aa})}{\overline{w}},$$

and the iterative equation for the expected frequency of allele $A$ (equivalent to the frequency of $A$ among gametes) at the next generation is

$$p_A^{(t+1)} = \frac{p_A^{(t)}(p_A^{(t)} w_{AA} + (1 - p_A^{(t)})w_{Aa})}{w_{AA}(p_A^{(t)})^2 + w_{Aa}2p_A^{(t)}(1 - p_A^{(t)}) + w_{aa}(1 - p_A^{(t)})^2}. \qquad (5.1)$$

| Genotype | $A/A$ | $A/a$ | $a/a$ |
|---|---|---|---|
| Fitness | $w_{AA}$ | $w_{Aa}$ | $w_{aa}$ |
| Frequency (at conception) | $p_A^2$ | $2p_A(1-p_A)$ | $(1-p_A)^2$ |
| Frequency (after selection) | $p_A^2 w_{AA}/\overline{w}$ | $2p_A(1-p_A)w_{Aa}/\overline{w}$ | $(1-p_A)^2 w_{aa}/\overline{w}$ |

Table 5.1 *Parameters of a biallelic single locus deterministic selection model. Mating is at random, so the genotype frequencies at conception are in Hardy-Weinberg equilibrium proportions (row 2 above ). The proportion of genotypes $A/A$, $A/a$ and $a/a$ that survive selection are given by $w_{AA}$, $w_{Aa}$ and $w_{aa}$, respectively. The relative proportions of genotypes after selection are given in row 3, where $\overline{w}$ is the mean fitness of the population (e.g., the proportion of the population that survives selection).*

To study the equilibrium allele frequencies under this model, we first derive a formula for the expected change in allele frequency per generation,

$$\begin{aligned}\Delta p_A &= p_A^{(t+1)} - p_A^{(t)} \\ &= \frac{p_A^{(t)}(1-p_A^{(t)})[p_A^{(t)}(w_{AA}-w_{Aa}) + (1-p_A^{(t)})(w_{Aa}-w_{aa})]}{w_{AA}(p_A^{(t)})^2 + w_{Aa}2p_A^{(t)}(1-p_A^{(t)}) + w_{aa}(1-p_A^{(t)})^2}.\end{aligned}$$

Clearly, $\Delta p_A = 0$ if $p_A = 0$ or $p_A = 1$. If **directional selection** is operating in favor of allele $A$, namely $w_{AA} \geq w_{Aa} > w_{aa}$, then the equilibrium frequency of allele $A$ is $p_A* = 1$. Conversely, if directional selection is operating in favor of allele $a$, namely $w_{aa} \geq w_{Aa} > w_{AA}$, then the equilibrium frequency of allele $A$ is $p_A* = 0$. It is also clear from the above equation that a third equilibrium exists (e.g., $\Delta p_A = 0$) if

$$p_A^{(t)}(w_{AA} - w_{Aa}) = -(1-p_A^{(t)})(w_{Aa} - w_{aa}).$$

Solving this equality for the equilibrium frequency of allele $A$ under this condition gives,

$$p_A* = \frac{w_{Aa} - w_{aa}}{2w_{Aa} - w_{aa} - w_{AA}}.$$

This is the equilibrium frequency under **overdominant selection** when the fitness of heterozgotes is greater than that of either homozygote,

namely $w_{Aa} > w_{AA}$ and $w_{Aa} > w_{aa}$. As an example of overdominance, let $w_{AA} = w_{aa} = \alpha$ and $w_{Aa} > \alpha$. Then at equilibrium,

$$\begin{aligned} p_A* &= \frac{w_{Aa} - w_{aa}}{2w_{Aa} - w_{aa} - w_{AA}} \\ &= \frac{w_{Aa} - \alpha}{2w_{Aa} - \alpha - \alpha} \\ &= \frac{w_{Aa} - \alpha}{2(w_{Aa} - \alpha)} = 1/2. \end{aligned}$$

**Sickle cell trait in Africa**

In sub-saharan Africa the frequency, $p_S$, of the mutant beta globin allele $\beta^S$ is between $0.10 \leq p_S \leq 0.20$. What is the relative fitness, $R = w_{SA}/w_{AA}$ of $\beta^S\beta^A$ genotypes versus $\beta^A\beta^A$, where $\beta^A$ denotes the non-mutant allele. Assume that the fitness of mutant homozygotes $\beta^S\beta^S$ (who develop sickle-cell anemia) is near zero ($w_{SS} = 0$) which would likely have been the case historically. Assume that the frequency of $\beta^S$ is at equilibrium so that,

$$p_S* = \frac{w_{SA} - w_{AA}}{2w_{SA} - w_{AA} - w_{SS}} = \frac{w_{SA} - w_{AA}}{2w_{SA} - w_{AA}} = \frac{R-1}{2R-1}.$$

Solving for $R$ as a function of $p*$ gives,

$$R = \frac{p*-1}{2p*-1} = \frac{1-p*}{1-2p*}.$$

If $p* = 0.10$ this gives,

$$R = 1 - 0.101 - 2(0.1) = 1.125,$$

so heterozygotes have a 12.5% increase in fitness, while if $p* = 0.20$ this gives,

$$R = \frac{1-0.20}{1-2(0.20)} = 1.333,$$

so heterozygotes have a 33% fitness increase. Note that if $w_{ss} > 0$ then $R$ would be smaller.

### 5.1.5 Genetic drift

Genetic drift is the term used to refer to the random changes in allele frequency that occur by chance sampling effects in small populations.

Genetic drift is an example of a "stochastic" process; if the drift process were repeated multiple times starting with the same initial population allele frequency the outcome would be different each time. The expected, or average, frequency at generation $t+1$, given the current frequency at generation $t$ is

$$\mathbb{E}(p^{(t+1)}|p^{(t)}) = p^{(t)}.$$

Thus, if we repeat the genetic drift process many times, starting with the same initial allele frequency $p^{(t)}$, the average frequency over all the repetitions is $p^{(t)}$, so the genetic drift process has no tendency to either increase, or decrease, allele frequency on average. However, in any particular realization of a population experiencing genetic drift the allele frequency will change in a random fashion. The variance of allele frequency at generation $t+1$ given the current frequency at generation $t$ is

$$var(p^{(t+1)}|p^{(t)}) = \frac{p^{(t)}(1-p^{(t)})}{2N},$$

where $N$ is the diploid population size. The greater the variance, the more random change occurs from one population to the next. The influence of genetic drift in changing allele frequencies decreases with an increase in the population size. In large populations, such as the current human population genetic drift has little influence relative to forces such as migration and selection. If an allele has current frequency $p^{(t)}$ then if no mutation is occurring, the probability that the allele is ultimately lost (e.g., $p^{(\infty)} = 0$) is $1 - p^{(t)}$ and the probability that it is ultimately fixed (e.g., $p^{(\infty)} = 1$) is $p^{(t)}$. If mutation is operating then an allele that is lost can be reintroduced by mutation so permanent fixation, or loss, does not occur.

# 7
# DNA Fingerprinting

## 7.1 Introduction

The concept of DNA fingerprinting originated in the early 1980s and is largely due to the work of Alec Jeffreys. He used a combination of restriction enzymes for cutting DNA into fragments based on the presence or absence of restriction sites and migration of DNA fragments on polyacrylamide gels followed by a Southern blotting (DNA-DNA hybridization) experiment using a radioactive probe to detect repetitive DNA fragments of variable length among individuals, referred to as minisatellites. A minisatellite is comprised of a core motif tens to hundreds of nucleotides in length that is repeated multiple times in a tandem array. If "core" denotes the core motif then an allele with $n$ copies is represented as $(\text{core})^n$. In the hybridization experiment the fragmented single-stranded DNA is probed with a radioactively labelled sequence bearing a core motif which labels all the alleles (at many locations in the genome) that represent variable repeat numbers of the motif. The multilocus profile of bands produced by Jeffreys' technique are ordered according to their size (migration distances in the gel). These profiles are expected to match between the DNA sample from a crime scene and that from a suspect in the case that the suspect is the source of the crime sample. Otherwise, they are expected to differ. Ignoring genotyping errors, different DNA profiles between suspect and crime scene sample will lead to the exclusion of the suspect.

Given the type of data produced by the Jeffreys' assay it is essentially impossible to calculate the probability of a random match for an individual that is not the perpetrator because it is not known which bands on a gel correspond to alleles at each locus. It is not even known how many loci are surveyed! Apart from this, there are other drawbacks to

Jeffreys' method: it requires a large amount of DNA (this technique was developed in the pre-PCR era) and it requires the use of a radioactive labelling procedure that is expensive and dangerous.

An alternative approach for DNA fingerprinting using PCR amplification of shorter DNA repeats, called short tandem repeats (STRs), or microsatellite loci, using locus-specific PCR primers, was developed in the late 1980s. The amplified DNA is run on a polyacrylamide gel and alleles corresponding to different numbers of repeats are identified using standard DNA staining techniques. No radioactive probe blotting is necessary making the procedure simple and safe. Also, the PCR step allows DNA fingerprinting using minute quantities of DNA from a crime scene. Microsatellite markers have now become the standard tool for DNA fingerprinting. An advantage of this approach that the alleles at each locus are unambiguously identified allowing population genetic procedures to be applied to calculate random match probabilities, etc. In addition to forensic applications, DNA fingerprinting is used in many other areas, including paternity analysis, and in wildlife genetics to carry out "genetic" mark-recapture studies to infer population size, etc.

## 7.2 DNA forensics

In the US, since the early 1990s a standard set of 13 microsatellite markers at unlinked autosomal loci plus two markers on the sex chromosomes (one on the X and one on the Y) have been used for DNA fingerprinting. These loci are referred to as the "Combined DNA Index System" or CODIS. The FBI maintains a large database of DNA fingerprint profiles (over 5 million), mostly of convicted felons, which was formally authorized by the DNA Identification Act passed in 1994. Because the number of CODIS loci is relatively small the probability of a random match cannot be neglected. This is especially true if a crime sample is compared with all 5 million individuals in the database of convicted felons. Thus, we need to use probability models to calculate the probability of chance matches. If this probability is very low there is stronger evidence that the suspect is the source of DNA from the crime scene when there is a perfect match of DNA fingerprint profiles.

### 7.2.1 Calculating DNA fingerprint match probabilities

We will begin by formulating the problem of interpreting DNA fingerprints from a crime scene sample and a suspect as a pair of mutually exclusive possibilities, stated as hypotheses to be tested:

- $H_p$: suspect left the DNA sample at the crime scene
- $H_d$: some other person left the DNA sample at the crime scene

The data are the multilocus genotype of the suspect, $G_S$, and the multilocus genotype from the crime scene DNA sample, $G_C$. In addition, there is other non-genetic evidence (witnesses to the crime, etc) denoted as $I$. The likelihood ratio will be used to evaluate the evidence and is defined as

$$LR = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)},$$

where $\Pr(X|Y)$ should be read as "the probability of $X$ given a fixed value of $Y$." This is called a conditional probability. If we assume that there is no genotyping error, then if the DNA genotypes from the suspect and the crime scene do not match,

$$\Pr(G_C|G_S, H_p, I) = 0 \text{ and } LR = 0.$$

In other words, the probability of observing the mismatched genotype from the crime scene sample given that the suspect is the perpetrator (e.g., given $H_p$ is true) is 0. So, this result produces a likelihood ratio of 0 and excludes the suspect. That simple.

If the DNA fingerprint from the crime scene matches that of the suspect, $G_C = G_S$, the situation is more complicated to evaluate. We cannot simply assume that the suspect is the source of the crime scene sample because it is always possible that more than one individual in a population has a given DNA fingerprint profile. Again, considering the LR we see that the numerator is,

$$\Pr(G_C|G_S, H_p, I) = 1.$$

In other words, if the suspect is indeed the perpetrator the crime scene sample DNA fingerprint must match the suspect's DNA fingerprint (e.g., with probability 1 they are identical) and so the numerator of the LR becomes 1. However, the denominator (which is the probability of a match given that the crime scene sample comes from another individ-

ual) is not zero and remains to be determined so that,

$$LR = \frac{1}{\Pr(G_C|G_S, H_d, I)}.$$

If uncertainty about $G_C$ is not influenced by $G_S$ (this may not be true, for example if the crime scene sample comes from a close relative of the suspect) then

$$LR = \frac{1}{\Pr(G_C|H_d, I)}.$$

Note that to calculate the probability of observing the genotype $G_C$ in an individual that is not the source of the crime scene DNA we need to know the population allele frequencies. As well, to predict genotype proportions from population allele frequencies, we will need to assume random mating (and HWE). Thus, we need to know population allele frequencies for a population that is "representative" of the population of which the suspect is a member. Population allele frequencies vary extensively among human ethnic groups and geographical regions and obtaining representative allele frequencies for such calculations is often problematic.

We now consider a simple example calculation using a single genetic locus for fingerprinting. Suppose that both the crime sample and the suspect are homozygous for allele A at the locus and therefore "match." Thus, $G_S = G_C = A/A$, and the frequency of $A$ in the population is $p_A = 0.8$. The likelihood ratio is

$$LR = \frac{1}{\Pr(A/A|p_A)} = \frac{1}{0.8^2} = 1.5625.$$

Larger values of the LR indicates increasing evidence that the suspect left the crime sample (e.g., supports $H_p$ versus $H_d$). A LRT of this magnitude would not convict a suspect. Increasing the number of loci surveyed will increase the significance of a match. Suppose that 3 loci are instead genotyped and we again obtain a perfect match. The genotypes are $G_C = G_S = A/A, B/b, c/c$ and the population allele frequencies are $p_A = 0.8$, $p_B = 0.7$, and $p_c = 0.5$. The loci are unlinked (and genotypes are therefore independent) and so the denominator of the likelihood ratio is a product of the probability for each individual locus,

$$LR = \frac{1}{\Pr(A/A|p_A) \times \Pr(B/b|p_B) \times \Pr(c/c|p_c)}.$$

Assuming HWE at each locus the probability is calculated as

$$LR = \frac{1}{0.8^2 \times [2(0.7)0.3] \times 0.5^2} = 14.88.$$

Thus, the evidence against the suspect is not an order or magnitude greater. For a sample of $L$ genetic loci, the general formula for the likelihood ratio, given a match between suspect and crime scene DNA fingerprints, is

$$1/LR = \prod_{i=1}^{L} \Pr(G_C(i)|p_i),$$

where $G_C(i)$ is the genotype at the $i$th locus and $p_i$ is the population allele frequency at locus $i$. We again assume HWE equilibrium to calculate the genotype probabilities given the allele frequencies at each locus. Note we have used the shorthand symbol,

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n.$$

earlier we noted that we needed to know the population allele frequencies to calculate random match probabilities. If an inappropriate population sample is used to estimate these frequencies (e.g., one that is not representative of the population of which the suspect is a member) this can lead to false incrimination.

To provide an example, suppose that we survey two genetic loci. We sample individuals from population 1, with allele frequencies $p_A = 0.01$ and $p_B = 0.01$ at each locus, to estimate allele frequencies. The suspect actually comes from population 2 in which the allele frequencies are $q_A = 0.99$ and $q_B = 0.99$. Suppose that the suspect is a match with genotypes A/A and B/B. Using allele frequencies from population 1, we calculate the probability of a random match as,

$$\Pr(A/A, B/B|p_A, p_B) = 0.01^2 \times 0.01^2 = 1/100,000.$$

However, if we had used the correct population frequencies (those of population 1) we would calculate the match probability as

$$\Pr(A/A, B/B|q_A, q_B) = 0.99^2 \times 0.99^2 = 0.9801$$

Thus, using the wrong population we have fairly damning evidence (LR = 100,000) while using the correct population there is no convincing evidence (LR = 1.02). This is an extreme example, but for diverse human

populations allele frequencies often vary by 30% or more, so it is not entirely unrealistic.

The bottom line in DNA forensics is that more genetic loci are always better, both in terms of reducing the risk of false incrimination and in increasing the evidence against the perpetrator. If modern genotyping technologies were used to genotype say 100,000 SNP loci, then every individual on earth would be utterly unique (apart from monozygotic twins) and we would not need to calculate probabilities of random matches: one would be virtually certain that that the sample came from the suspect in the case of a match (apart from cross-contamination) as well as being certain that it did not in the case of a mismatch (apart from genotyping errors). Unfortunately, the FBI and local law enforcement agencies have invested a large amount of effort in creating a huge database of felons using only 13 loci and so the fairly unreliable current DNA fingerprinting technologies persist. The concern about random matches is particularly significant when blind scanning a database of millions of samples using only 13 loci; in this case random matches may become uncomfortably probable.

# 8

# Parentage Analysis

DNA fingerprint profiles are also used for parentage analysis (most often paternity testing). Let $G_M$ be the genotype of the mother and $G_C$ be the genotype of the child. The genotype of the alleged father is $G_{AF}$. There are two mutually exclusive hypotheses to be tested. The first hypothesis is $H_p$: the alleged father is the biological father of the child. The second hypothesis is $H_d$: some other man is the father of the child. The genetic evidence is $E = \{G_M, G_C, G_{AF}\}$. We will consider two measures of support for paternity of the alleged father. The paternity index (PI) is a likelihood ratio (LR) of the probability of the genetic data given that the alleged father is the biological father and considering other non-genetic evidence, $I$, to the probability of the genetic data given that the father is some other man and considering $I$,

$$LR = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}.$$

The probability of paternity given the genetic and non-genetic evidence is defined as

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I) \times \Pr(H_p|I)}{\Pr(E|I)}.$$

This is obtained by applying Bayes' theorem and is sometimes referred to as the "posterior" probability of paternity. The denominator in this equation is 1 (e.g., $\Pr(E|I) = 1$) because the genetic data is assumed to be independent of other evidence of paternity. The first term in the numerator is the likelihood (the probability of the observed genetic data under the hypothesis of paternity and given the other evidence (which usually does not influence the genetics). The second term in the numerator is the "prior" probability of paternity based on other sources of

non-genetic evidence concerning possible paternity, $I$. The "posterior odds" of paternity is the ratio of the posterior probability of paternity versus non-paternity and simplifies to be the LR multiplied by the ratio of prior probabilities of paternity versus non-paternity,

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = LR \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}.$$

By some tedious algebra one can show that the formula for the posterior probability of paternity simplifies to

$$\Pr(H_p|E,I) = \frac{LR \times \Pr(H_p|I)}{LR \times \Pr(H_p|I) + [1 - \Pr(H_p|I)]}.$$

If there is no substantial prior evidence of paternity, the prior odds ratio is taken to be 1, so that $\Pr(H_d|I) = \Pr(H_p|I) = 0.5$. In that case, the probability of paternity simplifies to

$$\Pr(H_p|E,I) = \frac{LR \times 0.5}{LR \times 0.5 + [1 - 0.5]} = \frac{LR}{LR + 1}.$$

Thus, when the LR (PI) is small the probability of paternity is near zero and when it is large the probability of paternity approaches 1.

| $\Pr(H_p\|I)$ | LR(PI) | | | |
|---|---|---|---|---|
| | 1 | 10 | 100 | 1000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.001 | 0.001 | 0.010 | 0.091 | 0.500 |
| 0.500 | 0.500 | 0.910 | 0.990 | 0.999 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 8.1 *Posterior probability of paternity calculated for various values of the prior probability of paternity (rows) and the likelihood ratio (LR), also called the paternity index (PI) (columns).*

Table 8.1 shows the posterior probability of paternity as a function of prior probability of paternity (based on non-genetic evidence, $I$) and LR (or PI).

## 8.1 Calculation of the likelihood ratio

The likelihood ratio is defined to be the probability of the genotype data (for child, mother and alleged father) under the hypothesis that

the alleged father is the biological father, $H_p$, divided by the probability under the hypothesis that some other man is the father, $H_d$. This formula simplifies as follows,

$$\begin{aligned} LR &= \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}, \\ &= \frac{\Pr(G_C, G_M, G_{AF}|H_p, I)}{\Pr(G_C, G_M, G_{AF}|H_d, I)}, \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p, I)}{\Pr(G_C|G_M, G_{AF}, H_d, I)} \times \frac{\Pr(G_M, G_{AF}|H_p, I)}{\Pr(G_M, G_{AF}|H_d, I)}, \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p, I)}{\Pr(G_C|G_M, G_{AF}, H_d, I)} \times 1, \\ &= \frac{\Pr(G_C|G_M, G_{AF}, H_p, I)}{\Pr(G_C|G_M, G_{AF}, H_d, I)}, \end{aligned}$$

where the second term on the third line becomes 1 because the probabilities of genotypes for the mother and alleged father do not depend on whether the alleged father is the biological father (i.e., they only depend on population genotype frequencies). To calculate the numerator of the $LR$, we specify the maternal and paternal alleles received by the child, designated as $A_M$ and $A_F$, respectively, so that $G_C = (A_M, A_F)$. The probability of the child's genotype is then

$$\Pr(G_C|G_M, G_{AF}, H_p) = \Pr(A_M|G_M) \times \Pr(A_F|G_{AF}, H_p).$$

In general, we do not know whic of the child's alleles are maternally and paternally derived. If the child is homozygous at the locus, we do not need to know. For example, let $G_C = (A_i, A_i)$. Then the probability of the child's genotype under hypothesis $H_p$ is

$$\Pr(G_C|G_M, G_{AF}, H_p) = \Pr(A_M = A_i|G_M) \times \Pr(A_F = A_i|G_{AF}, H_p).$$

If the child is heterozygous at a locus, we must allow for both possible sources of alleles. For example, let $G_C = (A_i, A_j)$. Then the probability of the child's genotype under hypothesis $H_p$ is

$$\begin{aligned} \Pr(G_C|G_M, G_{AF}, H_p) = &\Pr(A_M = A_i|G_M) \times \Pr(A_F = A_j|G_{AF}, H_p) \\ &+ \Pr(A_M = A_j|G_M) \times \Pr(A_F = A_i|G_{AF}, H_p). \end{aligned}$$

As an example, suppose that $G_C = (A_1, A_2)$, $G_M = (A_1, A_2)$ and $G_{AF} = (A_1, A_1)$. By applying Mendel's law of independent assortment we have,

$$\Pr(A_M = A_1|G_M = (A_1, A_2)) = 0.5,$$

$$\Pr(A_M = A_2|G_M = (A_1, A_2)) = 0.5,$$
$$\Pr(A_F = A_1|G_{AF} = (A_1, A_1)) = 1,$$
$$\Pr(A_F = A_2|G_{AF} = (A_1, A_1)) = 0.$$

In words, the heterozygous mother is equally likely to transmit either of her two alleles, $A_1$ or $A_2$, while the homozygous father can only have transmitted allele $A_1$. The probability in the numerator for this example is then,

$$\begin{aligned}&\Pr(G_C = (A_1, A_2)|G_M = (A_1, A_2), G_{AF} = (A_1, A_1), H_p) = \\ &\Pr(A_M = A_2|G_M = (A_1, A_2)) \times \Pr(A_F = A_1|G_{AF} = (A_1, A_1)) \\ &+\Pr(A_M = A_1|G_M = (A_1, A_2)) \times \Pr(A_F = A_2|G_{AF} = (A_1, A_1)) \\ &= (0.5 \times 1) + (0.5 \times 0) = 0.5. \end{aligned} \tag{8.1}$$

To calculate the denominator, we assume that the father is a male that is randomly drawn from a population with allele frequencies $p = (p_{A_1}, p_{A_2})$. In that case,

$$\begin{aligned}&\Pr(G_C = (A_1, A_2)|G_M = (A_1, A_2), p, H_d) = \\ &\Pr(A_M = A_2|G_M = (A_1, A_2)) \times p_{A_1} \\ &+\Pr(A_M = A_1|G_M = (A_1, A_2)) \times p_{A_2} \\ &= \frac{1}{2}(p_{A_1} + p_{A_2}). \end{aligned} \tag{8.2}$$

The $LR$ is then

$$\frac{0.5}{0.5 \times (p_{A_1} + p_{A_2})} = 1/(p_{A_1} + p_{A_2}).$$

The LR approaches 1 if either of the alleles $A_1$ or $A_2$ become common (e.g., $LR \to 1$ as either $p_{A_1} \to 1$ or $p_{A_2} \to 1$). Intuitively, as $p_{A_1} \to 1$, it becomes more likely that the mother transmitted allele $A_2$ (which is rare in the population) and one of the common homozygous $(A_1, A_1)$ or heterozygous $(A_1, .)$ males in the population transmitted allele $A_1$ and so it becomes as likely that another man (other than $AF$) was the father. As $p_{A_2} \to 1$ individuals that carry allele $A_1$ become rare and so it becomes as likely that the mother transmitted allele $A_1$ and one of the common homozygous $A_2, A_2)$, or heterozygous $(A_2, .)$ males transmitted allele $A_2$. In fact, in this example if there are only two alleles in the population, $A_1$ and $A_2$, then $LR = 1$ regardless of the population allele frequencies (because $p_{A_1} + p_{A_2} = 1$). If there is a third allele, then the $LR$ can become large as the frequency of the third allele increases and the frequencies of alleles $A_1$ and $A_2$ both become small.

## 8.2 Parentage probabilities using multiple loci

As in the case of DNA fingerprinting, we assume that marker loci are unlinked (this is the case for the 13 marker loci in the CODIS database which are often used for parentage analysis as well). Assuming no linkage, we can calculate the $LR_i$ separately for each locus $i$ and then multiply to obtain the final LR,

$$LR = \prod_{i=1}^{L} LR_i,$$

where there are $L$ loci. To illustrate, we consider the simple case of $L = 2$ loci. Suppose that a child, mother and alleged father have the following 2 locus genotypes (where $A$ denotes the first locus and $B$ denotes the second locus),

$$\begin{aligned} G_C &= \{(A_1, A_2), (B_1, B_1)\}, \\ G_M &= \{(A_1, A_1), (B_1, B_2)\}, \\ G_{AF} &= \{(A_1, A_2), (B_1, B_1)\}, \end{aligned}$$

and let the population allele frequencies be,

$$\begin{aligned} p_{A_1} &= 0.7, \\ p_{A_2} &= 0.3, \\ p_{B_1} &= 0.8, \\ p_{B_2} &= 0.2. \end{aligned}$$

For locus $A$, the numerator is $1/2$ and the denominator is $p_{A_2}$, so $LR_A = 1/(2p_{A_2})$. For locus $B$, the numerator is $1/2$ and the denominator is $p_{B_1}/2$, so $LR_B = 1/p_{B_1}$. Thus the LR is,

$$LR = LR_A \times LR_B = \frac{1}{2p_{A_2}} \times \frac{1}{p_{B_1}} = \frac{1}{2p_{A_2}p_{B_1}} = \frac{1}{2 \times 0.3 \times 0.8} = 2.083.$$

If we assume that $\Pr(H_p|I) = \Pr(H_d|I) = 0.5$ then the posterior probability of paternity is

$$\Pr(H_p|E, I) = \frac{LR}{LR + 1} = \frac{2.083}{2.083 + 1} = 0.675.$$

Because alleles at both loci are relatively common there is not much evidence for paternity in this case. If the father carries rare alleles, the probability of paternity increases. For example, if the population allele

frequencies were instead

$$\begin{aligned} p_{A_1} &= 0.99, \\ p_{A_2} &= 0.01, \\ p_{B_1} &= 0.01, \\ p_{B_2} &= 0.99, \end{aligned}$$

then the $LR$ becomes

$$LR = \frac{1}{2 \times 0.01^2} = 5000,$$

and the probability of paternity becomes

$$\Pr(H_p|E, I) = \frac{5000}{5000 + 1} = 0.9998.$$

# 10
# DNA Substitution Models

The evolutionary history of humans and other species can be inferred by analyzing patterns of DNA nucleotide substitutions. The evolutionary history is represented as a phylogenetic tree describing the pattern of shared ancestry of species and levels of molecular divergence (branch lengths). If fossils are available to calibrate the ages of ancestors for some nodes of the phylogeny, a "molecular clock" may sometimes be used to infer the ages of other nodes without fossil calibrations. Here we introduce basic concepts for modeling DNA substitutions between species, estimating species divergence times and inferring phylogenetic trees from sequence data.

## 10.1 Alignment of homologous sequences

The basic premise of a phylogenetic analysis is that sequences are inherited with modification from a common ancestor. A particular site in a sequence is homologous in descendent species A and B if the site was inherited from the same site of the sequence in the common ancestor of A and B. If no changes occur in the descendent sequences alignment is easy – just match up the sequences so that all paired sites have identical nucleotides. Most often, changes will have occurred in the descendent sequences, making alignment challenging. Sequences in descendents may change due to mutation as well as insertion or delection of one or more nucleotides. The goal of sequence alignment is to align homologous sites in two or more sequences descended from a common ancestral sequence. Most alignment algorithms do this by allowing sequence "gaps" due to insertions or deletions (indels) as well as changes due to point mutation. Penalties are applied for mismatches (substitutions)

and for indels and a search is made for the alignment with the highest score (fewest penalties). In this Chapter we will assume that an alignment is available and will focus on inferring phylogeny, etc, using a set of aligned sequences.

## 10.2 Pairwise percentage of substitions

One simple measure of the level of molecular divergence between a pair of species is the percentage of sites in the aligned sequences that have different nucleotides, defined as

$$d = \frac{x}{n},$$

where $x$ is the number of sites at which the two species have different nucleotides and $n$ is the total number of sites in the sequence. A fixed difference of a nucleotide in a sequence between species is referred to as a DNA substitution. Note that a substitution occurs in two steps: first, a new mutation arises in one of the species; second, the mutation becomes fixed in the species due to processes such as genetic drift and natural selection.

If genetic drift is the only process operating (e.g., changes at a site are selectively neutral), Motoo Kimura showed that the rate of substitution is equal to the site-specific mutation rate. This can be derived as follows: at each generation, each diploid individual experiences a mutation at a particular site with rate $\mu$ per homologous chromosome and the expected number of new mutations in a diploid population of size $N$ is then $2N\mu$. Under genetic drift, Sewall Wright showed that the probability of fixation of an allele present in $i$ copies is $i/2N$. A newly arisen mutation is initially present as a single copy and therefore the probability of fixation is $1/2N$. Thus, the expected rate of substitution, $v$, per generation is

$$v = 2N\mu \times \frac{1}{2N} = \mu.$$

In many genomic regions, nucleotide changes are not neutral (particularly in coding regions, or regulatory regions) and the rate of substitution may be either increased (positive selection) or decreased (negative selection). For example, third codon positions are highly redundant (many changes in the nucleotide at the third codon position do

not change the amino acid that is coded for) and thus tend to experience higher substitution rates than first, or second, codon positions; this is because most proteins are under negative selection (e.g., changes to the protein amino acid sequence have negative effects on fitness).

## 10.3 Modeling DNA substitutions

The percentage of substitutions tends to underestimate the actual number of substitutions that occurred because if multiple substitutions occur at a particular site this can regenerate the ancestral nucleotide in the descendent making it appear as if no substitutions occurred. To deal with this problem, an explicit model of DNA substitution is needed. One of the earliest (and simplest) models of DNA substitution was developed by Jukes and Cantor in 1969. The JC69 model assumes that mutations occur according to a Poisson probability distribution. The probability that $M$ substitutions occur at a particular nucleotide site during a period of time, $t$, is then given by

$$\Pr(M) = \frac{e^{-\nu t}(\nu t)^M}{M!},$$

where $\nu$ is the rate of substitution per unit of time. The probability that no substitutions occur is then

$$\Pr(M = 0) = e^{-\nu t},$$

and the probability that one or more substitutions occur is

$$\Pr(M \geq 1) = 1 - \Pr(M = 0) = 1 - e^{-\nu t}.$$

The JC69 model further assumes that when a substitution occurs it results in a change to any of the 4 nucleotides with equal probability, $1/4$. Given this model, the probability that one or more substitutions occur at a site that change an ancestral allele T to an allele A in the descendent after time $t$ is

$$p_{TA}(t) = (1 - e^{-\nu t})\frac{1}{4}.$$

The first term of the above equation is the probability that one or more substitutions occur during time $t$. The second term is the probability that the last substitution to occur generated nucleotide A. The probability that the descendent has the ancestral allele T after time $t$ is

$$p_{TT}(t) = e^{-\nu t} + (1 - e^{\nu t})\frac{1}{4} = \frac{1}{4} + \frac{3}{4}e^{-\nu t}.$$

The first term of the above equation gives the probability that no substitutions occurred during time $t$ (in which case both the ancestral and descendent site have nucleotide T) and the second terms gives the probability that one or more substitutions occurred and the final substitution generated allele T.

## 10.4 Substitution proportions under JC69 model

We can determine the expected proportion of sites with substitutions between a pair of sequences from species that separated $t$ time units in the past as follows

$$p_{i \neq j}(t) = (1 - e^{-\nu t})\frac{3}{4}.$$

This is just the probability that one or more substitutions occur multiplied by the probability that the final substitution is to a nucleotide other than the ancestral nucleotide. This gives the probability of a substitution for a particular site; if sites are assumed to be independent then this also gives the expected proportion of sites with substitutions. Similarly, the expected proportion of sites that are identical is

$$p_{i=j}(t) = e^{-\nu t} + (1 - e^{\nu t})\frac{1}{4} = \frac{1}{4} + \frac{3}{4}e^{-\nu t}.$$

In the derivations above we allow a nucleotide to change to itself. Normally, a change to the same state would not be considered a substitution. To deal with this, we modify the substitution rate so that only a fraction $3/4$ of the changes are considered substitutions (e.g., those that result in a different nucleotide). Therefore, we define a "corrected" rate, $v$ as

$$v = \frac{3}{4}\nu.$$

Solving for $\nu$ as a function of $v$ gives

$$\nu = \frac{4}{3}v.$$

Replacing $\nu$ by $\frac{4}{3}v$ in all the previously derived equations generates the standard formulas for the JC69 model.

## 10.5 JC69 distance and divergence time

The proportion of nucleotide substitutions between a pair of sequences from different species can be used to estimate the divergence time if we assume that the substitution rate is identical in the two species (this is known as the molecular clock hypothesis). The total time separating the two sequences is twice the divergence time. Solving for $t$ in the JC69 formula for the expected proportion of substitutions gives,

$$\hat{t} = -\frac{1}{v}\left(\frac{3}{4}\right)\log(1 - 4/3\hat{p}),$$

where $\hat{p}$ is the observed proportion of substitutions between a pair of sequences. If the substitution rate, $v$, is known then the divergence time between the species can be estimated as $\hat{t}/2$. As an example, the proportion of substitutions in non-coding nuclear DNA sequences between human and chimpanzee is roughly 0.010 to 0.015. The rate of mutation in mammalian nuclear genes is roughly $v = 10^{-9}$ per year and assuming that the non-coding regions are neutral we can use the mutation rate as an estimate of the substitution rate. The predicted divergence time between human and chimpanzee is then

$$\frac{1}{2}\hat{t} = -\frac{1}{2} \times \frac{1}{10^{-9}}\left(\frac{3}{4}\right)\log(1 - 4/3 \times 0.01) = 5033633,$$

or about 5 MYA assuming $\hat{p} = 0.01$ or

$$\frac{1}{2}\hat{t} = -\frac{1}{2} \times \frac{1}{10^{-9}}\left(\frac{3}{4}\right)\log(1 - 4/3 \times 0.015) = 7576015,$$

or about 7.5 MYA assuming $\hat{p} = 0.015$.

## 10.6 Kimura 2 parameter model

The JC69 model is not very realistic for human sequence data because it assumes that transitions are as likely as transversions. Empirical evidence from comparisons among species (as well as experimental evidence from mutation studies) suggests that transitions are much more likely than transversions. In 1980, Kimura proposed a more realistic model of DNA substitution that allows for different rates of transition versus transversion. In the so-called Kimura 2 parameter model (or K80), there are two parameters. The first parameter is the relative

rate of transitions versus transversions,

$$\kappa = \frac{\alpha}{\beta},$$

where $\alpha$ is the rate of transitions and $\beta$ is the rate of transversions. The second parameter is the overall substitution rate $v$ defined as

$$v = \alpha + 2\beta.$$

Under this model, the probability of a transition is

$$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-4vt/(\kappa+2)} - \frac{1}{2}e^{-2vt(\kappa+1)/(\kappa+2)},$$

and the probability of a transversion is

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-4vt/(\kappa+2)}.$$

This formulation produces two equations (for $p_1(t)$ and $p_2(t)$) in two unknowns ($\kappa$ and $t$). Letting $S = p_1(t)$ and $V = 2p_2(t)$ be the expected proportions of transitions and transversions among a sample of independently evolving sites and solving for $\kappa$ and $t$ gives,

$$\hat{t} = -\frac{1}{v}\left(\frac{1}{2}\log(1-2S-V) - \frac{1}{4}\log(1-2V)\right),$$

and,

$$\hat{\kappa} = \frac{2\log(1-2S-V)}{\log(1-2V)} - 1.$$

# 14
# Linkage Analysis

Strictly speaking, Mendel's law of independent of assortment only applies to loci that are located on different chromosomes. However, the process of recombination during meiosis in humans and other species can create partial independence between marker loci located on the same chromosome. Early work in Drosophila and other species showed that the expected fraction of recombinant haplotypes depends on the physical distance separating loci. The greater the distance, the greater the fraction of recombinants. In humans, co-segregation of marker loci can be observed by genotyping individuals in pedigrees. These "linkage analysis" studies were used to create the first genetic maps (linkage maps) of the locations of markers on chromosomes. In the post-genomic era, linkage analysis has been used to study variation in recombination rates across the human genome and as a preliminary step for locating gene mutations that are a cause of disease. Here, we describe the basic elements of linkage analysis and provide some example applications to the estimation of recombination rates in humans, inference of linkage maps, and disease gene mapping.

## 14.1 Probability model of recombination

In humans, recombination occurs during meiosis at the stage after the maternal and paternal chromosomes have been replicated. A characteristic structure forms called a chiasma that is associated with a double stranded cleavage of DNA and exchange of segments between a pair of randomly chosen chromatids. Each replicated chromosome (or sister chromatid) participates in half the chiasma on average. Let $n$ be the number of chiasma formed on the interval. Let $r_n$ be the probability

that a gamete is recombinant if there are $n$ chiasma. Note that $r_0 = 0$. A recursive formula can be written for the probability that a gamete is recombinant given $n$ chiasma,

$$r_n = \frac{1}{2} r_{n-1} + \frac{1}{2}(1 - r_{n-1}),$$

where the first term of the sum is the probability that the gamete does not participate in the $n$th crossover multiplied by the probability that it is recombinant after $n - 1$ crossovers (in which case it is still recombinant), and the second term is the probability that the gamete does participate in the $n$th crossover multiplied by the probability that it was not recombinant after $n - 1$ crossovers (in which case it becomes recombinant). By direct substitution, we see that

$$r_1 = \frac{1}{2} r_0 + \frac{1}{2}(1 - r_0) = \frac{1}{2} 0 + \frac{1}{2}(1 - 0) = \frac{1}{2},$$

and

$$r_2 = \frac{1}{2} r_1 + \frac{1}{2}(1 - r_1) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{2},$$

and in general $r_n = 1/2$ if $n > 0$.

## 14.2 Haldane's map function

A useful measure of the distance between markers on a chromosome is the expected fraction of recombinant gametes, $\theta$. A simple model of recombination developed in the 1930s by J.B.S. Haldane assumes that double-stranded breaks (and chiasma) form with equal intensity along the interval between a pair of markers. The physical length of the interval, $d$, is assumed to be large and the rate of breaks (chaisma) is assumed to be small. This results in a Poisson distribution for the number of chiasma formed on the interval during meiosis. Let $\theta$ be the recombination fraction (per meiosis) between markers A and B. The expected fraction of recombinant gametes is,

$$\theta = \frac{1}{2} \Pr(n > 0).$$

Assuming that chiasma form with rate $c$ per Mb on the interval between a pair of markers, A and B, according to a Poisson distribution,

$$\Pr(n) = \frac{e^{-cd}(cd)^n}{n!},$$

where $d$ is the distance between A and B in units of Mb. For the Poisson distribution, we have

$$\Pr(n > 0) = 1 - \Pr(n = 0) = 1 - e^{-cd},$$

and therefore

$$\theta = \frac{1}{2}(1 - e^{-cd}). \tag{14.1}$$

Note that as $cd$ becomes small $\theta$ tends to zero and as $cd$ becomes large $\theta$ tends to $1/2$, these are the minimum and maximum observable recombinant fractions of gametes. If $\theta = 1/2$ we say that the loci are unlinked, while if $\theta = 0$ we say that they are in complete linkage.

## 14.3 Inferring recombination rates

By genotyping parents and children and observing the frequency of recombinant haplotypes in the children of particular pairs of parents we can use genotyped individuals from a pedigree to infer rates of recombination in a specific region of the genome. We first consider a simple estimator of recombination rate, $c$, using a so-called linear approximation. The exponential function can be represented as an infinite series (here we have used a Taylor series expanded about the point $x = 0$),

$$e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{24} - \cdots$$

If $x$ is small, terms such as $x^2$, $x^3$, etc, will be small and can be neglected, leading to the approximation,

$$e^{-x} \approx 1 - x.$$

Substituting this approximation for $e^{-x}$ into equation 14.1 above (with $x = cd$) gives,

$$\theta \approx \frac{1}{2}(1 - [1 - cd]) = \frac{1}{2}cd.$$

Solving the above equation for $c$ in terms of $\theta$ and $d$ gives the estimator,

$$\hat{c}_1 \approx \frac{2\theta}{d}$$

This estimator will only be accurate if $\theta \leq 0.01$. An estimator can also be obtained without the linear approximation by solving equation 14.1

for $c$ directly,

$$\hat{c}_2 = -\frac{1}{d}\log(1-2\theta).$$

The recombination fraction is usually measured in units of centimorgans (cM), named after the famous Drosophila geneticist Thomas Hunt Morgan. 1 cM equals 1% recombination (a fraction 0.01) per meiosis. The rate $c$ is often presented in units of cM/Mb.

As an example, consider a collection of 500 unrelated family trios (parents plus one child) for which the recombinant haplotypes of children in a region targeted for genotyping could be unambiguously identified. Let the number of recombinant maternally-derived haplotypes be $Y_m = 124$ and let the number of recombinant paternally-derived haplotypes be $Y_p = 86$. The total fraction of recombinants is

$$\theta = \frac{124+86}{1000} = 0.21.$$

Using the approximate estimator gives

$$\hat{c}_1 = \frac{2\times 0.21}{1.6} = 0.2625 = 26.25 \text{ cM/Mb}.$$

The exact estimator gives,

$$\hat{c}_2 = -\frac{1}{1.6}\log(1-2\times 0.21) = 0.340 = 34 \text{ cM/Mb}.$$

In this example, there is a considerable discrepancy between the approximate and exact estimators (this is expected since $\theta$ is large) and the exact estimator should be more accurate in this case. Note that in the above analysis, we have estimated the sex-averaged recombination rate. There is considerable variation between rates of recombination in males and females and they are often estimated separately.

## 14.4 Linkage maps

A linkage map arranges genetic markers of unknown location in a genome into ordered arrangements on chromosomes based on a comparison of relative rates of recombination (or observed propotions of recombinants). Less recombination between markers implies that they are physically closer to one another on a chromosome. One simple procedure is to use pairwise values of $\theta$ to order markers on a chromosome. For example, with three genetic markers, A, B and C, if the pairwise

recombination proportions are $\theta_{AB} = 0.01$, $\theta_{BC} = 0.05$ and $\theta_{AC} = 0.06$ the optimal ordering would be A-B-C.

## 14.5 Linkage mapping of disease genes

The objective of linkage mapping is to estimate the recombination proportion, $\theta$, between one or more genetic markers and an unobserved disease locus. Linkage mapping is most effective for so-called simple Mendelian genetic disorders. The recurrence of simple Mendelian disorders is completely determined by genotype. The "penetrance" of a genotype at a disease locus is defined as the probability of developing the disease given the genotype, $\Pr(\text{disease}|\text{genotype})$. For a disease locus with two classes of alleles (e.g., disease alleles, $D$, and normal alleles, $d$) there are three penetrance parameters,

$$\begin{aligned} f_1 &= \Pr(\text{disease}|DD) \\ f_2 &= \Pr(\text{disease}|Dd) \\ f_3 &= \Pr(\text{disease}|dd). \end{aligned}$$

For a simple recessive Mendelian disorder $f_1 = 1$, $f_2 = 0$ and $f_3 = 0$, while for a simple dominant disorder $f_1 = f_2 = 1$ and $f_3 = 0$. The parameter $f_3$ is referred to as the phenocopy rate (e.g., the rate at which individuals without any disease alleles develop the disease). Most simple Mendelian disorders such as cystic fibrosis (a recessive disorder) have a phenocopy rate of zero.

To develop the theory underlying linkage mapping of disease genes, we consider the specific case of a rare dominant disease caused by an allele $D$ at a disease locus, with $d$ to be the normal allele. We assume complete penetrance ($f_1 = f_2 = 1$) and no phenocopies ($f_3 = 0$). Because the disease allele is rare, we can assume that our families are comprised entirely of those where one parent is an affected individual with a heterozygous genotype at the disease locus, $Dd$, and the other is a non-affected individual with disease locus genotype $dd$. This is because if $p_D$ is the frequency of the disease allele, $D$, under HWE,

$$\begin{aligned} f(DD) &= p_D^2, \\ f(Dd) &= 2p_D(1-p_D), \\ f(Dd \times Dd) &= f(Dd)^2 = (2p_D(1-p_D)^2)^2, \\ f(Dd \times dd) &= f(Dd) \times f(dd) = p_D(1-p_D) \times (1-p_D)^2, \end{aligned}$$

where $f(Dd \times dd)$ denotes the frequency of matings between individuals with $Dd$ and $dd$, genotypes, etc. The frequencies $f(Dd)$ and $f(Dd \times dd)$ are proportional to the allele frequency $p_D$, whereas the other frequencies, $f(Dd \times Dd)$, etc, are proportional to $p_D^2$. Thus, for a rare disease allele with $p_D \to 0$ we can neglect all genotypes and matings other than $Dd$ and $Dd \times dd$.

The possible genotypes at a marker locus for trios of families with one affected parent that is heterozygous at the marker locus, $P_1 = (Dd, Aa)$, one unaffected parent that is homozygous at the marker locus, $P_2 = (dd, aa)$, and one child, $C$, (either affected, or unaffected) are $P_1 = Aa$, $P_2 = aa$, $C = Aa$ and $P_1 = Aa$, $P_2 = aa$, $C = aa$. Under the null hypothesis, marker allele A is linked to disease allele $D$ and marker allele $a$ is linked to disease allele $d$. The possible gametes produced by each parent (and their probabilities) are shown in Table 14.1. The probability for each of the 4 possible combinations of disease status

| Parent genotype | Gamete | Probability |
|---|---|---|
| Aa/Dd | A-D | $(1/2)(1-\theta)$ |
| Aa/Dd | A-d | $(1/2)\theta$ |
| Aa/Dd | a-D | $(1/2)\theta$ |
| Aa/Dd | a-d | $(1/2)(1-\theta)$ |
| aa/dd | a-d | 1 |

Table 14.1 *Probabilities of each possible gamete for parents with genotypes Aa/Dd and aa/dd assuming linkage of the marker to the disease locus with parameter $\theta$*

and marker genotype in children are given in Table 14.2 The transmis-

| Child genotype | Child disease status | Probability | No. Families |
|---|---|---|---|
| Aa | affected | $(1/2)(1-\theta)$ | $Y_1$ |
| Aa | normal | $(1/2)\theta$ | $Y_2$ |
| aa | affected | $(1/2)\theta$ | $Y_3$ |
| aa | normal | $(1/2)(1-\theta)$ | $Y_4$ |

Table 14.2 *Probabilities of each possible combination of marker genotype and disease status in a child born to parents with genotypes Aa/Dd and aa/dd assuming linkage of the marker to the disease locus with parameter $\theta$.*

sion of disease and marker alleles within families are independent and so the probabilities multiply to obtain the total probability of the ob-

served counts of families with each of the four possible configurations of marker genotype and disease status for the child,

$$\begin{aligned}\Pr(Y_1, Y_2, Y_3, Y_4|\theta) &= \prod_{i=1}^{4} \Pr(Y_i|\theta), \\ &= \left[\frac{1}{2}(1-\theta)\right]^{Y_1} \times \left[\frac{1}{2}\theta\right]^{Y_2} \times \left[\frac{1}{2}\theta\right]^{Y_3} \times \left[\frac{1}{2}(1-\theta)\right]^{Y_4}, \\ &= \left[\frac{1}{2}(1-\theta)\right]^{Y_1+Y_4} \left[\frac{1}{2}\theta\right]^{Y_2+Y_3}.\end{aligned}$$

To estimate $\theta$ by the method of maximum likelihood we search for a value of $\theta$ that maximizes the probability of the observed data as calculated above. By convention, the base 10 logarithm of the probability is maximized which is equivalent to maximizing the probability directly. The base 10 logarithm of the probability is called the lod score, $z(\theta)$. For the rare dominant disorder outlined above the lod score is,

$$z(\theta) = (Y_1 + Y_4)\log_{10}([1-\theta][1/2]) + (Y_2 + Y_3)\log_{10}(\theta[1/2]).$$

If the lod score is maximized at $\theta = 0.5$ this indicates no linkage of the marker locus to a disease locus. If it is maximized at $\theta = 0$ this indicates a very tight linkage.