Chapter 2
Topics in Quantitative Genetics

Abstract An understanding of the genetics of current world populations provides the conceptual basis upon which today's genetic association studies rest. This chapter focuses specifically on gaining a basic grounding in three general topics:

1. ***Linkage disequilibrium* (LD), the nonrandom associations of alleles**. A discussion on how linkage disequilibrium varies between populations due to multiple factors, such as random drift in allele frequencies in isolated populations, population migration, *admixture*, and population expansion.
2. **Population heterogeneity**. A discussion of the effects of population heterogeneity, including population stratification, admixture, and relatedness between subjects—specifically, the distribution of *marker alleles* and their apparent association with each other and with causal variants, using marker data for the empirical *estimation of relatedness* and *kinship coefficients* and of *identity by descent* probabilities.
3. **The common disease-common variant hypothesis**. Arguments in favor of the *common disease-common variant hypothesis* and a discussion of distributions of allele frequencies for both marker alleles and causal variants.

In addition to these concepts, several important tools for the investigation of linkage disequilibrium and of population heterogeneity are introduced: specifically data from the HapMap project and data manipulation and LD visualization tools which help to explore these data effectively. *Principal components analysis* (PCA) of large-scale genetic data for the purpose of examining population substructure and admixture is introduced and illustrated using a download of phase 3 HapMap data for 11 population samples; an example of some simple PLINK commands and a corresponding *R* script is provided. These illustrate the selection (in PLINK) of a subset of SNPs to be used in the computation and display, by *R*, of leading principal components that characterize global population structure.

D.O. Stram, *Design, Analysis, and Interpretation of Genome-Wide Association Scans*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-9443-0_2,

Many parts of the broader field of population genetics are entirely ignored here. Notably we do not discuss, except parenthetically, natural selection as a force determining allele frequency distributions within modern populations or the differences seen between modern populations. Differences between populations in allele frequencies for marker alleles or causal variants are largely assumed here to be due to random drift or founder effects and population expansion.

Implicitly we are restricting interest, and this is part of topic (3), to common genetic causes of disease. The mere fact that the alleles are common indicates that the reproductive fitness of carriers of the alleles is not greatly impacted; even when looking at *diseases* that do affect reproductive fitness (such as fatal childhood diseases, early adult onset mental illnesses), the selective pressure against alleles that cause modest increases in risk of such disease may be very minor if the diseases themselves are rare.

2.1 Distribution of a Single Diallelic Variant in a Randomly Mixing Population

2.1.1 Hardy–Weinberg Equilibrium

The description in Chap. 1 of the mechanics of gamete formation provides the basis for a discussion of first the marginal distribution of allele counts for a single individual and secondly the joint distribution of these counts between related individuals. The *Hardy–Weinberg rule* states that under a number of assumptions, the marginal distribution of the number of copies of a given allele observed in a single individual will follow a binomial distribution with index (i.e., the "number of trials") equal to 2 and mean parameter equal to the frequency in the population of that allele. For example, if we have a diallelic marker taking the values A with frequency p and a with frequency $(1 - p)$, then the number of copies, n_A, of A that a given individual carries will take the values 0, 1, and 2 with probabilities equal to $(1 - p)^2$, $2p(1 - p)$, and p^2, respectively. If an allele count, n_A, follows this distribution, then we say that it is in *Hardy–Weinberg equilibrium* with the name being reflective of the independent derivation of this rule by Godfrey Hardy and Wilhelm Weinberg in 1908. This marginal distribution applies to all individuals sampled from a given population for which there has been at least one generation of random mating; additional assumptions are that there is no reproductive disadvantage to carrying one or two of the two alleles (e.g., the allele does not increase the risk of early mortality and infertility) and that the number of copies carried by males is the same as that carried by females, e.g., the rule does not apply to X chromosome variants in males. Violations of random mating include the interrelated concepts of inbreeding, population stratification, and admixture, each of which will be discussed in the following.

2.1.2 Random Samples of Unrelated Individuals

Large-scale genotyping technology frees genetic studies from needing closely related individuals with disease, allowing for the (generally much easier) sampling of unrelated subjects from within a given population. If we have a sample of size N from a population in which the allele of interest is in Hardy–Weinberg equilibrium, and if the N individuals sampled are unrelated to each other (which is a reasonable expectation so long as the number of individuals sampled is small relative to the size of the population being targeted), then the distribution of the sum $\sum_{i=1}^{N} n_{iA}$ of the allele counts of allele A for each individual i will follow a binomial distribution with index equal to $2N$ and frequency also equal to p. In this case we can estimate p as $\hat{p} = \frac{1}{2N}\sum_{i=1}^{N} n_{iA}$. The variance of this estimator is then equal to

$$\frac{1}{(2N)^2}\sum_{i=1}^{N}\mathrm{Var}(n_{iA}) = \frac{1}{(2N)^2}2Np(1-p) = \frac{1}{2N}p(1-p).$$

For values of N and p which jointly meet the requirement that $\min(2Np, 2N(1-p))$ is "large," the estimator $\hat{p}$ can be approximated as a normal random variable with mean p and standard error equal to $\sqrt{\frac{1}{2N}\hat{p}\,(1-\hat{p}\,)}$ so that an approximate $1-\alpha$ level confidence interval for p is given by

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{2N}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{2N}}. \tag{2.1}$$

The accuracy of such a confidence interval depends importantly on the type I error rate α; for the "traditional" α level of 0.05 , a value of $2Np$ of around 5 or so is generally considered adequate for the approximation to be reliable; however, for more rigorous α levels such as those often employed in GWAS studies (e.g., $\alpha = 5 \times 10^{-8}$), $2Np$ should be considerably larger (at least five times as large for common alleles) before approximate confidence intervals (and equivalently tests) can be safely relied upon. See Chap. 7 for more information.

2.1.3 Joint Distribution Between Relatives of Allele Counts for a Single SNP

In the following we consider the joint distribution of the number of copies, n_{iA}, and n_{jA} of the same diallelic variant taking values a and A, for two related individuals drawn from a randomly mixing population. Since the population is randomly

Table 2.1 All possible transmissions and number of copies of chromosomal region shared identically by descent for two full siblings

	Sibling 2			
Sibling 1	MGM, PGM	MGM, PGF	MGF, PGM	MGF, PGF
MGM, PGM	2	1	1	0
MGM, PGF	1	2	0	1
MGF, PGM	1	0	2	1
MGF, PGF	0	1	1	2

mixing, the marginal distribution of both n_{iA} and n_{jA} will follow the Hardy–Weinberg equilibrium, with the same value of allele frequency p, but the probabilities that each take the values 0, 1, or 2 will not be independent of each other. For example, if we select a mother and one of her children, the child will always share exactly one of the alleles of the mother, and so if the mother has genotype *AA* (for example), then it is impossible (in the absence of a de novo mutation) that the child will have n_{jA} equal to 0. This is because one chromosome is always passed from each parent to each offspring.

2.1.3.1 Identity by Descent

If we assume that all mating is between unrelated individuals, i.e., that the population is "outbred," then the degree of relatedness between individuals for whom a pedigree structure is known can be summarized in terms of the expected numbers of alleles that are inherited identically by descent from the founders of the pedigree. In order to clarify what inheritance *identical by descent* (*IBD*) means, consider a simple nuclear family with two parents and two offspring and we will label the four founders (parents) copies of a short chromosomal region according to whether the copy originated from each of the maternal grandmother (MGM), maternal grandfather (MGF), paternal grandmother (PGM), or paternal grandfather (PGF). Table 2.1 considers all 16 (equally probable for neutral alleles) combinations that could be transmitted to the two offspring and gives the number of copies of this segment that are therefore shared identically by descent. Since each cell has probability 1/16, we see that the probability (call it z_0) that the two siblings share no alleles is 1/4; the probability, z_2, they share two alleles is also 1/4; and the probability, z_1, that they share 1 allele is 1/2. Thus the expected number of shared copies is equal to $0 \times z_0 + 1 \times z_1 + 2 \times z_2 = 1$; dividing this by two gives the expected fraction of shared alleles as $(1/2)z_1 + z_2 = (1/2)$. Similar calculations for parent-offspring pairs give the three IBD sharing probabilities as $z_0 = 0$, $z_1 = 1$, and $z_2 = 0$ with the fraction of shared alleles again equal to (1/2).[1]

[1] If however the maternal and/or paternal grandparents were related to each other, then there is a nonzero probability that a parent–offspring pair will share 2 alleles IBD, i.e., $z_2 > 0$. For example, if the parents are siblings, then the three IBD probabilities will be $z_0 = 0$, $z_1 = 5/8$, and $z_2 = 3/8$; *see below*.

Table 2.2 Joint probability distribution for counts, n, of number of copies of allele A given number of alleles shared IBD

			Number of alleles shared IBD		
Unphased genotype	n_1	n_2	0	1	2
(aa, aa)	0	0	$(1-p)^4$	$(1-p)^3$	$(1-p)^2$
(Aa, aa)	1	0	$2p(1-p)^3$	$p(1-p)^2$	0
(AA, aa)	2	0	$p^2(1-p)^2$	0	0
(aa, Aa)	0	1	$2p(1-p)^3$	$p(1-p)^2$	0
(Aa, Aa)	1	1	$4p^2(1-p)^2$	$p(1-p)^2 + (1-p)p^2$	$2p(1-p)$
(AA, Aa)	2	1	$2p^3(1-p)$	$p^2(1-p)$	0
(aa, AA)	0	2	$p^2(1-p)^2$	0	0
(Aa, AA)	1	2	$2p^3(1-p)$	$p^2(1-p)$	0
(AA, AA)	2	2	p^4	p^3	p^2

Such calculations can be extended to more complex pedigrees, and it can readily be shown, for example, that both half siblings and grandparent–grandchild pairs are expected to share 1/4 of their alleles IBD.

Now let us consider the *correlation* between the counts, $n_{A,1}$ and $n_{A,2}$, of an SNP allele observed in two related individuals (numbers 1 and 2) with a known relationship. We can compute the covariance between $n_{A,1}$ and $n_{A,2}$ as a function of the probabilities (z_0, z_1, z_2) of sharing either 0, 1, or 2 alleles identically by descent. To show this, we first need to describe the joint probability distribution of the count of a marker allele conditional on IBD status.

Table 2.2 shows the joint probabilities of n_1 and n_2 conditional on the number of alleles shared identically by descent. These probabilities are computed by simple counting assuming that both the shared and unshared alleles are sampled independently from a population of possible alleles. For example, the conditional probability of the genotype (Aa, Aa) being sampled given that 1 allele is shared in common can be broken into a fraction with numerator equal to

$$\begin{aligned}
&\Pr\left(Aa, Aa \text{ is sampled} \middle| \text{1st and 3rd alleles shared IBD}\right) \times \Pr\left(\text{1st and 3rd alleles shared IBD}\right) \\
&+ \Pr\left(Aa, Aa \text{ is sampled} \middle| \text{1st and 4th alleles shared IBD}\right) \times \Pr\left(\text{1st and 4th alleles shared IBD}\right) \\
&+ \Pr\left(Aa, Aa \text{ is sampled} \middle| \text{2nd and 3rd alleles shared IBD}\right) \times \Pr\left(\text{2nd and 3rd alleles shared IBD}\right) \\
&+ \Pr\left(Aa, Aa \text{ is sampled} \middle| \text{2nd and 4th alleles shared IBD}\right) \times \Pr\left(\text{2nd and 4th alleles shared IBD}\right)
\end{aligned}$$

and denominator

$$\begin{aligned}
&\Pr(\text{1st and 3rd alleles shared}) + \Pr(\text{1st and 4th alleles shared}) \\
&+ \Pr(\text{2nd and 3rd alleles shared}) + \Pr(\text{2nd and 4th alleles shared}).
\end{aligned}$$

Consider the first probability Pr(*Aa*, *Aa* is sampled|1st and 3rd alleles shared). Since genotypes *Aa* and *aA* are treated here as indistinguishable, this probability can be further broken into the probability of sampling first an *A* (for the first and third alleles), then an *a* (for the 2nd allele), and then another *a* for the fourth allele, plus the probability of sampling first an *a* (for the first and third alleles) and then two *A* s in sequence (for the 2nd and 4th alleles). Thus, this probability is equal to $p(1-p)^2 + (1-p)p^2$. The same quantity is found for the 3 other sampling probabilities, and moreover the probability of sharing either the 1st and 3rd or 1st and 4th or 2nd and 3rd or 2nd and 4th are all equal to each other, i.e., are equal to some constant, *c*. Therefore the total conditional probability is

$$\frac{4c\left[p(1-p)^2 + (1-p)p^2\right]}{4c} = p(1-p)^2 + (1-p)p^2,$$

as presented in the table.

If we know (from the pedigree relationship) the probabilities, z_0, z_1, and z_2, of sharing zero one or two alleles IBD, then multiplying the rightmost three columns of Table 2.2 by z_0, z_1, or z_2, respectively, and summing them together, we have the complete joint probability distribution of n_1 and n_2. Hence, we can compute the covariance of n_1 and n_2 as a simple weighted sum of the possible values of $[n_1 - E(n_1)][n_2 - E(n_2)] = (n_1 - 2p)(n_2 - 2p)$ with the weights given by the joint probability distribution for n_1 and n_2. A little bit of algebra shows that

$$\mathrm{Cov}(n_1, n_2) = (z_1 + 2z_2)p(1-p).$$

Since the variance of n_{jA} is equal to $2p(1-p)$ for both $j = 1$ and $j = 2$, we see that the correlation between the allele counts for the two individuals will be $\mathrm{Cor}(n_1, n_2) = (1/2)z_1 + z_2$ which is from above the expected fraction of shared alleles. From this we can write the $n \times n$ covariance matrix of a single SNP allele for all related subjects, as equal to

$$\mathrm{Var}(n_1, n_2, \ldots, n_N) = 2p(1-p)\mathbf{K}, \tag{2.2}$$

where the matrix **K** has diagonal elements equal to 1 and off-diagonal elements equal to $k_{ij} = \langle (1/2)z_1 + z_2 \rangle_{ij}$ with z_1 and z_2 computed for each pair (i, j) of family members.

2.1.4 Coefficients of Kinship and of Inbreeding

Up until now the *relationship matrix* **K** has been a correlation matrix with diagonal elements all equal to 1. **K**, however, does not have to be a correlation matrix and off-diagonal elements can in fact be greater than one. Note that $z_1 + 2z_2$ is the expected number of alleles shared identically by descent between

two individuals. Now consider the following experiment: randomly sample two chromosomal segments, one from each individual, and then count the fraction of times that the two alleles are identical. It is easy to see that on average this fraction will (since there are four possible ways of sampling the two alleles) be equal to 1/4 times the expected number of shared alleles. It is this fraction, i.e., $(1/4)z_1 + (1/2)z_2$ that is called the *coefficient of kinship*. Now think of a single individual and consider two independent draws from the same individual's chromosomes; if that individual's parents are unrelated, the two chromosomal segments of interest that comprise the two alleles are different, and the probability that the same one is sampled twice is 1/2. If however the parents are related with a coefficient of kinship equal to h, then the probability that the same chromosome is represented twice in that individual, rather than just once, will be equal to h. Thus when sampling two alleles at random with replacement from the same individual, the probability that the same allele is sampled twice is $(1/2) \times (1 - h) + 1 \times h$ or $(1/2)(1 + h)$. Moreover, it can readily be shown that the variance of the number of alleles carried by a single individual is $2p(1 - p)(1 + h)$. Here, h (i.e., the kinship between parents) is termed the *coefficient of inbreeding* and applies to individuals and identical twins. Thus for inbred populations, we continue to have the model for the variance-covariance matrix of a single marker measured for a total of N individuals equal to

$$2p(1-p)\mathbf{K}, \tag{2.3}$$

but now with $\mathbf{K}$ being equal to twice the *kinship matrix* which we denote with the bold Greek lowercase letter kappa, $\boldsymbol{\kappa}$. The kinship matrix $\boldsymbol{\kappa} = (1/2)\mathbf{K}$ has off-diagonal terms equal to $\langle (1/4)z_1 + (1/2)z_2 \rangle_{ij}$ for $i \neq j$ and diagonal terms equal to $(1/2)(1 + h_i)$.

2.2 Relationship Between Identity by State and Identity by Descent for a Single Diallelic Marker

Identity by state (IBS) refers to the number of similar alleles shared (irrespective of descent) between two individuals at a particular locus. For example, if one subject has genotype *aa* and another subject genotype *Aa*, then one allele (the *a*) is said to be identical by state; if both subjects have genotypes equal to *Aa* and *Aa*, then two alleles are identical by state. From the joint distribution, in Table 2.2, of genotypes for two subjects given IBD status, we can easily compute the conditional probability distribution for sharing 0, 1, or 2 alleles by state, given the number of alleles that are identical by descent. This is shown in Table 2.3.

Table 2.3 The probability of allele sharing (IBS) given identity by descent (IBD) for a diallelic marker with allele frequency p

		Number of alleles shared IBD		
		0	1	2
Pr(IBS|IBD)	0	$2p^2(1-p)^2$	0	0
	1	$4p(1-p)^3 + 4p^3(1-p)$	$2p(1-p)^2 + 2p^2(1-p)$	0
	2	$4p^2(1-p)^2 + (1-p)^4 + p^4$	$(1-p)^3 + p(1-p)^2 + p^2(1-p) + p^3$	1

2.3 Estimating IBD Probabilities from Genotype Data

Table 2.3 motivates a simple method of moments estimate of IBD probabilities using a set of M markers, under the assumptions that subjects are drawn from a single population; improvements on this method to deal with finite sample sizes in estimation of allele frequency is given in [1] and maximum likelihood estimation is described in [2].

Changing notation to add an index ℓ distinguishing markers, we can estimate z_0 by counting the number of alleles ($\ell = 1, \ldots, M$) with IBS count equal to zero. Since from Table 2.3 the marginal probability, Pr $(IBS(n_{i\ell}, n_{i'\ell}) = 0)$, of zero IBS sharing is equal to $z_0\{2p_\ell{}^2(1-p_\ell)^2\}$ so that the expected number of alleles with IBS $= 0$ is $2z_0\sum_{\ell=1}^{M} p_\ell{}^2(1-p_\ell)^2$. This suggests that we can estimate z_0 by equating the observed count, $\sum_{\ell=1}^{M} I\{\text{IBS}(n_{i\ell}, n_{i'\ell}) = 0\}$, of alleles with IBS $= 0$ with its expectation $2z_0\sum_{\ell=1}^{M} p_\ell{}^2(1-p_\ell)^2$ and solve for z_0 as

$$\hat{z}_0 = \frac{\sum_{\ell=1}^{M} I\{\text{IBS}(n_{i\ell}, n_{i'\ell}) = 0\}}{2\sum_{\ell=1}^{M} p_\ell{}^2(1-p_\ell)^2}. \tag{2.4}$$

Similarly, since the marginal probability that $\text{IBS}\left(n_{i\ell}, n_{i'\ell}\right) = 1$ is equal to $z_0\{4p_\ell(1-p_\ell)^3 + 4p_\ell{}^3(1-p_\ell)\} + z_1\{2p_\ell(1-p_\ell)^2 + 2p_\ell{}^2(1-p_\ell)\}$, we can estimate z_1 as

$$\hat{z}_1 = \frac{\sum_{\ell=1}^{M} I\{\text{IBS}(n_{i\ell}, n_{i'\ell}) = 1\} - \hat{z}_0\sum_{\ell=1}^{M} 4p_\ell(1-p_\ell)^3 + 4p_\ell{}^3(1-p_\ell)}{\sum_{\ell=1}^{M} 2p_\ell(1-p_\ell)^2 + 2p_\ell^2(1-p_\ell)} \tag{2.5}$$

and finally estimate z_2 as $\hat{z}_2 = 1 - \hat{z}_0 - \hat{z}_1$.

We must also estimate the allele frequencies, p_ℓ as well generally using only unrelated subjects, and these frequencies are assumed known in the above. Note that the estimators for the IBD probabilities are only valid for members of homogeneous populations and do not apply when one or more non-mixing hidden strata exist. Applying these formulas to stratified samples overestimates z_2, the probability of sharing two alleles IBD, since a hallmark of population stratification is an overrepresentation of homozygotes (relative to that expected under HWE) for markers that are differentiated between two or more groups (see below).

2.4 The Covariance Matrix for a Single Allele in Nonrandomly Mixing Populations

We have just shown that when sampling related individuals, the covariance matrix for a single allele with frequency p is equal to $2p(1-p)\mathbf{K}$ with the off-diagonal elements of $\mathbf{K}$ reflecting the relationships between subjects and the diagonals reflecting inbreeding. It turns out that for *structured populations*, i.e., ones where there are either hidden non-mixing populations and/or incomplete admixture between such groups, a similar model holds, except that in these cases the off-diagonal elements of $\mathbf{K}$ are not directly reflective of familial relationships between individuals but are rather influenced by the similarities or differences in genetic ancestry between individuals who would not normally be considered to be related (i.e., randomly selected members of a large racial/ethnic group) compared to other subjects (i.e., those from other such groups). The important thing (as shown for several examples below) is that in structured populations, the relationship matrix (or more properly *quasi-relationship* matrix since the elements of the matrix are not reflective of close familial relationships but rather of population history and ancestry) is the same for all variants (assuming neutral effects).

2.4.1 Hidden Structure and Correlation

Consider the between-person correlation in the allele counts for a given marker that is induced by the presence of hidden structure in a population. Intuitively it is clear that the individuals who are more alike (i.e., members of the same group) will have more similar values of a marker that has different allele frequencies in the different groups than do individuals who are in different groups. In order to quantify this and to make some general observations, we use a well-known model (the *Balding-Nichols* beta-binomial model) for the difference in allele frequencies in currently isolated groups which have originated from the same ancestral group, with the difference in allele frequencies being due to random drift in the frequencies through time. While a simplification of population genetics models for drift in allele frequencies [3], the model is used extensively [4].

The Balding-Nichols [5] beta-binomial model assumes that for any marker allele with frequency equal to p in the ancestral population, the allele frequency p_t in a modern population (here t indexes the different populations) will be distributed according to the beta distribution

$$p_t \sim B\left(\frac{1-F}{F}p, \frac{1-F}{F}(1-p)\right), \tag{2.6}$$

so that p_t will have the same mean p as in the ancestral population and variance equal to $Fp(1-p)$. Here F is a parameter that is common for all SNPs and specifies the genetic distance, due to random drift in neutral allele frequency, between the modern day and ancestral populations. We assume that given the allele frequency, an individual's marker genotypes in the modern population (or populations) will then be distributed as draws from a binomial distribution—i.e., will satisfy Hardy–Weinberg equilibrium.

Now assume that a study sample is made up of subjects from total of T different subpopulations each with a genetic distance from the ancestral population equal to F_t and consider first single individuals and then pairs of subjects from the same or different populations. Again let n_i be the count of the number of copies of a given SNP for the person i in the sample, who happens to be in subpopulation t. We now compute the mean and variance of n_i unconditionally, i.e., incorporating the variability of the modern day allele frequency. The expected value of n_i can be computed as $E_{p_t}\left[E\left(n_i|p_t\right)\right] = E_{p_t}(2p_t) = 2p$, just as in the ancestral population. The variance of n_i is computed as

$$\mathrm{Var}(n_i) = E_{p_t}\left[\mathrm{Var}\left(n_i|p_t\right)\right] + \mathrm{Var}_{p_t}\left[E\left(n_i|p_t\right)\right].$$

With

$$\begin{aligned} E_{p_t}\left[\mathrm{Var}\left(\mathrm{n_i} \mid \mathrm{p_t}\right)\right] &= E_{p_t}[2p_t(1-p_t)] = 2E_{p_t}(p_t) - 2E_{p_t}\left(p_t^2\right) \\ &= 2p - 2\left(p^2 + F_t p(1-p)\right) \end{aligned}$$

and

$$\mathrm{Var}_{p_t}\left[E\left(n_i|p_t\right)\right] = \mathrm{Var}_{p_t}(2p) = 4F_t p(1-p),$$

so that the unconditional variance of n_i equals $2p(1-p)(1+F_t)$. Notice that this is overdispersed relative to the binomial variance and implies that counts from a structured population will not follow the Hardy–Weinberg rule (and there will be an over abundance of homozygotes and a corresponding deficit in the number of heterozygotes compared to that expected under HWE). Of course if we knew which population individual i was sampled from, then we could compute the conditional variance, $2p_t(1-p_t)$ which does correspond to HWE, but here we are assuming that this is not possible.

Now suppose that two individuals i and j are sampled from the same subpopulations. Analogous to the rule for variances we have

$$\mathrm{Cov}(n_i, n_j) = E_{p_t}\left[\mathrm{Cov}(n_i, n_j|p_t)\right] + \mathrm{Cov}_{p_t}\left[E(n_i|p_t), E(n_j|p_t)\right].$$

The first term is zero for all p_t (since we assume independence within each population). The second term is $\mathrm{Cov}(2p_t, 2p_t)$, i.e., equals the variance of $2p_t$, which is $4F_t p(1-p)$. Thus the covariance between the two genotypes is $4F_t p(1-p)$ and the correlation between the two genotypes is $2\frac{F_t}{(1+F_t)}$. For individuals in different subpopulations, we easily see that the covariance and correlation between n_1 and n_2 is zero.

Note (Homework) that under this model the covariance matrix of the sum, s, of two independent alleles with ancestral allele frequency p_1 and p_2 is equal to $(2p_1(1-p_1) + 2p_2(1-p_2))\mathbf{K}$ and hence that the correlation of a sum of alleles between two individuals in the same population will also be equal to $2\frac{F_t}{(1+F_t)}$ just as for a single allele. This extends to *polygenes* composed of weighted sums of many different alleles.

2.4.1.1 Relationship Between Balding–Nichols' F Parameter and the Fixation Index F_{st}

Measures of the degree of population stratification in a given population include the fixation index F_{st} described initially by Sewall Wright [6] which quantifies degree of population separation by the difference in heterozygote frequency expected in stratified versus randomly mixing populations. This statistic can be written as $F_{st} = \frac{H_t - H_s}{H_t}$. Here H_t is the expected fraction of heterozygotes in the total stratified population if HWE could be assumed in that population, which we specify as $2\overline{p}(1-\overline{p})$, and H_s is the average of the expected fraction of heterozygotes within each component of the stratified population, namely, $E(2p_t(1-p_t))$. Now consider generating a SNP with ancestral frequency p for T equal-sized subpopulations using the beta-binomial model with all subpopulations sharing the same F. It is easy to see from the calculations immediately above that the expected number of heterozygotes in the full stratified population will be $E(2p_t(1-p_t)) = 2p(1-p)(1-F)$. If we know the ancestral allele frequency, p, then we can say that the expected fraction of heterozygotes in the stratified population under HWE is $2p(1-p)$, so that $F_{st} = F$. If we do not know the ancestral allele frequency, then we estimate $\overline{p}$ as 1/2 the observed count of the generated allele over all populations. The expected value of $2\overline{p}(1-\overline{p})$ (i.e., H_t) under the beta-binomial model can be shown to be equal to $H_t = 2p(1-p)(1-F/T)$ so that the calculated value of F_{st} is approximately $\frac{T-1}{T-F}F$ which converges to F as T (the number of populations) increases. See the simulation experiment in file *BN_Fst.r*. Note that we do not (in the simulation) calculate a separate F_{st} for each SNP and then average the F_{st}. Instead we calculate H_t and H_s for each SNP, then average these values over all the SNPs and form the ratio from the average values to estimate F_{st}.

2.4.2 *Effects of Incomplete Admixture on the Covariance Matrix of a Single Variant*

Population admixture occurs when individuals from two or more previously separated populations begin interbreeding. As mentioned earlier, even one generation of subsequent random mixing leads to alleles with a marginal distribution that are in Hardy–Weinberg equilibrium no matter how distinct the merging populations are. However, modern day admixture is more complicated than this, typical admixed populations show a greater than expected variation between members of the current population in the number of ancestors that derive from each of the mixing populations (compared to random mating after an initial mixing), and this between-person variation in ancestry is what leads to a failure of the Hardy–Weinberg rule to apply.

Consider two populations mixing, then as described by Chen et al. [7] the covariance matrix for a single SNP is of form $2p(1-p)\mathbf{K}$ where the elements, k_{ij}, depend upon the fraction of ancestors α_i and α_j, that each subject has from one of the two mixing populations (with $1-\alpha_i$ and $1-\alpha_j$ from the other population). Here we are assuming that each sampled subject is unrelated to each other in the usual sense.

In this case the diagonal terms, k_{ii} of $\mathbf{K}$ will equal $1+F_i$ with $F_i = F(1+2\alpha_i^2 - 2\alpha_i)$ and off-diagonal terms $F_{ij} = F(1+2\alpha_i\alpha_j - \alpha_i - \alpha_j)$. This reduces to the above model when $\alpha_i = 1$ and $\alpha_j = 0$ or vice versa. A subtlety noted by Chen et al. is that when all α_i are equal (i.e., there is complete mixing), the effects of admixture are no longer detectable, and replacing p from the ancestral population with the allele frequency in the completely admixed population (which is now $p_1\alpha + p_2(1-\alpha)$ where p_1 and p_2 are the allele frequencies in the two mixing populations) allows one to drop the F_i and F_{ij} in the above calculations, i.e., the admixed population can be treated as any other homogeneous population when considered alone, at least in terms of the correlation of a single SNP.

2.5 Direct Estimation of Differentiation Parameter *F* from Genotype Data

We now consider the problem of using genotype data from two individuals to estimate a pair-specific differentiation parameter F. We concentrate here on the problem of there being hidden non-mixing populations, indexed by t. For simplicity we assume that all F_t are equal to a single value of F.

Using the index ℓ for distinguishing different markers from each other, note that the variances of the allele counts, $n_{i\ell}$, for each marker ℓ all have the same form, i.e., the variance is equal to $(1+F)$ times the usual binomial variance $2p_\ell(1-p_\ell)$ where p_ℓ is the allele frequency in the ancestral population. Similarly the covariance between variants for two individuals in population l is always equal to a constant, $2F$, multiplied by the usual binomial variance. While it is impossible to estimate the (unknown) differentiation parameter F between subjects based on just

one marker observed in the two individuals, in a GWAS study we have hundreds of thousands of SNP markers available. Suppose for the time being that the ancestral allele frequency p_ℓ of marker ℓ is known. Then upon observing a total of M SNPs, we could form standardized alleles for each SNP and each individual, i, as

$$z_{i\ell} = \frac{n_{i\ell} - 2p_\ell}{\sqrt{2p_\ell(1-p_\ell)}}. \tag{2.7}$$

Each of these has expected value 0 and variance $(1 + F)$. Moreover the covariance between two different standardized alleles, $z_{i\ell}$ and $z_{j\ell}$, for the same SNP but different individuals is equal to

$$\frac{\mathrm{Cov}(n_{i\ell}, n_{j\ell})}{2p_\ell(1-p_\ell)} = \frac{4Fp_\ell(1-p_\ell)}{2p_\ell(1-p_\ell)} = 2F.$$

This leads to the estimator of the covariance matrix of the counts of a given marker for individuals i and j as

$$2p_\ell(1-p_\ell)\hat{\mathbf{K}}, \tag{2.8}$$

where $\hat{\mathbf{K}}$ is a 2×2 matrix with off-diagonal elements equal to

$$\frac{1}{M}\sum_{\ell=1}^{M} z_{i\ell}z_{j\ell} \tag{2.9}$$

and the ith diagonal element equal to

$$\frac{1}{M}\sum_{\ell=1}^{M} z_{i\ell}^2. \tag{2.10}$$

Extending this to consider all pairs of individuals, the resulting matrix is commonly used for describing the structure of a population sample and is the matrix computed, for example, in the principal components approach described by [4] and in the EIGENSTRAT program.

Although we have described it in terms of simple hidden structure problem (non-mixing hidden groups), this estimator is useful in many problems where we can model the covariance matrix for all SNPs as equal to a constant matrix $\mathbf{K}$ times a variance parameter unique to each SNP (here $2p_\ell(1 - p_\ell)$).

2.5.1 Relatedness Revisited

We have seen from our initial discussion of relatedness that the covariance matrix for the marker counts, $n_{i\ell}$ and $n_{j\ell}$, for two related subjects, i and j, coming from an otherwise unstratified population (i.e., randomly mixing) is equal to

$$2p_\ell(1-p_\ell)\begin{pmatrix} 1 & \frac{1}{2}z_1+z_2 \\ \frac{1}{2}z_1+z_2 & 1 \end{pmatrix},$$

with z_0 and z_0 the IBD probabilities as defined above so that our estimate $\hat{\mathbf{K}}$ adapts to this setting as well.

If there is no hidden population structure, but we relax the assumption that all subjects are unrelated, then $\hat{\mathbf{K}}$ estimates the *relationship matrix* (i.e., twice the kinship matrix $\boldsymbol{\kappa}$ [8]) which has off-diagonal elements equal to $(1/2)z_1 + z_2$ and diagonal elements equal to $1 + h_i$ where h_i is the inbreeding coefficient for subject i, i.e., the kinship between subject i's parents. In the forgoing we refer to $\hat{\mathbf{K}}$ (computed now for all pairs of subject i and j) as the estimated relationship matrix, recognizing, however, that this name is only technically accurate in the absence of population stratification.

Note that if one was sure that the underlying population was not stratified, then we could also use (2.4) for $\hat{z}_0$ and (2.5) for $\hat{z}_1$ (and $\hat{z}_2 = 1 - \hat{z}_0\text{-}\hat{z}_1$) to estimate $(1/2)$ $z_1 + z_2$, and this estimator may be better behaved when using markers with small allele frequencies p_k because division is delayed until after considerable simulation has taken place. Recently Yang et al. [9] have specifically noted that the diagonal estimate $\frac{1}{M}\sum_{\ell=1}^{M} z_{i\ell}^2$ is poorly behaved if many markers with small allele frequencies are used and have suggested the alternative

$$1+\frac{1}{M}\sum_{\ell=1}^{M}\frac{n_{i\ell}^2-(1+2p_\ell)n_{i\ell}+2p_\ell^2}{2p_\ell(1-p_\ell)},$$

which has the same expectation but a smaller sampling variance than does $\frac{1}{M}\sum_{\ell=1}^{M} z_{i\ell}^2$.

Even in situations where hidden stratification is present, it still is of interest to identify close relatives in a study especially in the process of performing data cleaning and outlier identification (see later chapters). One simple method of estimating IBD probabilities for close relationships that is not overly sensitive to population stratification is to simply exclude from the calculations of relatedness any markers that appear to be out of Hardy–Weinberg equilibrium. More complete estimation of z_1 and z_2 in a fashion that is relatively impervious to hidden structure is described in Purcell et al. [1] in their discussion of the GWAS analysis program, PLINK. The general approach is to bound the estimates of z_1 and z_2 so that they correspond to actual familial relationships. In almost all cases (except for identical twins, duplicated DNA samples, or extremely inbred population), the true probability of sharing two alleles IBD will be less than the probability of sharing one

allele IBD, so that z_2 is generally much closer to zero than is z_1; in randomly mixing populations, z_2 can be no greater than the square of twice the kinship coefficient. Enforcing such constraints on z_2 helps to better estimate true close relationships in the presence of hidden structure.

As shown above for two individuals sampled from an admixed population derived from the (incomplete) mixing of two different ancestral populations, the covariance matrix for a given marker will again equal a constant matrix **K**, with **K** not depending on which SNP is measured, multiplied by a variance parameter depending on SNP [7]. Even more general problems, such as admixture + relatedness or inbreeding within pedigree founders in a complex pedigree, can be modeled in this fashion [8, 10].

2.5.2 Estimation of Allele Frequencies

Next we need to point out that for our hidden subpopulation example that we have assumed that the ancestral population frequency is known when calculating the quasi-relationship matrix $\hat{\mathbf{K}}$. This is rarely if ever true in actual practice. What is done is to estimate p_ℓ as the frequency of marker ℓ within the sample that is available. In the hidden strata example, the overall sample allele frequency remains an unbiased estimator of p_ℓ, but it is not one with particularly good statistical properties. For example, it will not in general converge to the true ancestral p_ℓ as the number, N, of subjects in the study increases (only if the number of distinct hidden strata all derived from the same ancestral population increases with N would the sample allele frequency be expected to converge to p_ℓ). Rather than seeking to improve the estimate of p_ℓ, it makes better sense in most cases to simply accept the fact that $\hat{\mathbf{K}}$ measures the relative relatedness between subjects rather an absolute relatedness [10]. In fact if we use the sample estimated allele frequency in the calculations, then we will see that kinship estimates for certain pairs of subjects are negative reflecting that these pairs are more dissimilar to each other than are pairs from the same population. While some authors recommend setting negative off-diagonal elements to zero, this is not usually necessary for the kind of uses we make of this estimator [10] (as in Chap. 4).

2.6 Allele Frequency Distributions

2.6.1 Initial Mutations and Common Ancestors

One of the underlying principles of population genetics relevant to all association studies is that, at least in most cases, all carriers of a given genetic variant inherited that variant from a single most recent common ancestor (MRCA); i.e., we can trace

back from generation to generation the variant allele to a single carrier chromosome or, more precisely, a single chromosomal segment, or *locus*, surrounding the variant allele. Indeed it is this "single origin" of a given variant that partly leads to the concept of association-based genetic studies, since the background pattern of other variants or genetic markers, and specifically SNPs, that were already present on the ancestral segment uniquely labels that given chromosome (distinguishing it from all other chromosomes with progeny in today's population). Searching for unique patterns of SNPs that are related to disease by association with causal alleles is at the core of what is meant by association studies.

This most recent common ancestral chromosome was not necessarily the earliest chromosome that ever carried that genetic variant, for at the time that the MCRA was extant, there may have been many other carriers of that same variant allele. Indeed the variant allele may already have been common in the population at that time. In essence, however, no other carrier of that allele at that time left progeny in today's population. As we see shortly, this is an overstatement, since a different genetic variant, for example, one on a different chromosome, will have a different most recent ancestral origin. It is not that the other carriers of the first allele have no progeny today; they just have no progeny today *at that locus*. It is through the process of recombination that different segments of the DNA have different origins.

This tracing back of carriers of a variant lying on a particular chromosomal locus or to its common origin can be modeled in terms of a tree-diagram called the coalescent. Coalescent theory has many important results, the simplest of which can be derived easily under restricted assumptions such as no recombination, no natural selection, random mating, and fixed population size, yielding essentially the Wright-Fisher model of neutral evolution [11].

Under the assumption of neutral evolution, if we consider a population of chromosomes which has been of constant size N over its history, we can construct a genealogy for that population by thinking of each offspring chromosome present at generation t as "randomly choosing" its parent chromosome from the population (also of size N) of chromosomes present at time $t - 1$, and from this we can derive the probability distribution of the time (counting backwards in generations) until any two chromosomal segments selected from among the current generation meet their MCRA or *coalesce*.

The probability that these two segments coalesce, i.e., choose the same parent, in the previous generation is $1/N$, while the probability that they do not coalesce is $1 - 1/N$. Repeating this process through earlier generations we see that the probability distribution of t the time (in generations counting backwards $t = 1, 2, \ldots$) until coalescent occurs is geometrically distributed with probability distribution function

$$P(t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}.$$

The first term indicates the probability that the coalescence did not occur at times $t = 1, 2, \ldots, t-1$, and the second term is the probability of a coalescence at time t. If N is large, this can be approximated as the exponential distribution

$$P(t) = \frac{1}{N} e^{-\frac{t}{N}},$$

which has mean N and variance equal to N^2.

More generally if we start with a sample of n chromosomal segments from among the population of size N, then we see that the time to the first coalescent (from among ${}_nC_2 = \binom{n}{2}$ possible pairs) will be distributed as

$$P_n(t) = (1 - p_n)^{t-1} p_n,$$

where p_n is the probability of at least one coalescent occurring in one generation.

If N is large relative to n, then we can approximate p_n as

$$\frac{{}_nC_2}{N}$$

and also ignore the possibility that more one coalescent occurs in a single generation. Thus under the coalescent approximation, the number of distinct lineages decreases in steps of one back in time and the expected time from the kth coalescent until the $k + 1$st is equal to

$$\frac{N}{{}_{n-k}C_2}. \tag{2.11}$$

The expected time to the MCRA for the whole sample will be equal to the sum of the coalescent times or

$$N \sum_{k=0}^{n-2} 1/{}_{n-k}C_2 = 2N(1 - 1/n). \tag{2.12}$$

We can also (for large N) approximate $P_{n-k}(t)$, the distribution of time between coalescent F and $k + 1$, as exponential with mean $\frac{N}{{}_{n-k}C_2}$.

This exponential approximation is heavily exploited for computer simulation of realistic genetic data [12]. For example, under the assumptions above, plus an assumption about the probability (per generation) of a new (neutral) mutation, we can take a population of chromosomal segments and simulate the occurrence of new genetic markers (such as SNPs or any other inheritable variant) on those segments by doing simulations backward in time, rather than forward. Moreover the program does not have to keep track of whether coalescences occur in each generation,

instead it generates $N - 1$ independent exponential random numbers, with the appropriate mean, and independently of these a random bifurcating tree to construct the full genealogy describing the relatedness between all chromosomes in the current generation.

2.6.2 Mutations and the Coalescent

A basic simplifying assumption often useful in population genetics is that mutations only occur once at a given site with no possibility of backwards mutation or further changes at that site. This assumption is known as the infinite sites model, i.e., that there are an infinite number of mutation sites so that the chance of two mutations occurring at exactly the same site can be neglected. This model reflects the large number of base pairs in the genome as well as a very low per-base pair (10^{-8} or so) probability of change per generation or so in "higher organisms" [13]; the model appears to apply to SNPs considering, for example, the very small fraction of SNPs that have more than two alleles.

In coalescence models mutations are assigned along the branches of the tree, with exponential arrival times so that the number of new mutations along a particular path is distributed as a Poisson random variable with mean equal to the product of the mutation rate and the length of that path. To accommodate this notion in the simulation, each mutation is assigned not only a path position but also a unique chromosomal position.

Simulation programs based on the coalescent have wide usefulness in illuminating and comparing the properties of various statistical methods when they are to be applied to real data. Coalescent-based simulation programs, such as those given by Hudson [12] while based on this fundamental approach, can be modified to allow certain of the assumptions of the Wright-Fisher model to be relaxed, in particular population sizes can be allowed to change in time, and the migration of genes between populations can be accommodated.

See Fig. 2.1 for a depiction of a simulated coalescent tree.

2.6.3 Allelic Distribution of Genetic Variants

An important topic that impacts the design, analysis, and prospect for success of GWAS studies is the frequency distribution of variation in the genome and specifically of variation that is causative of disease or influences other phenotypes of interest. Much of the motivation for undertaking first candidate gene association studies and later GWAS studies was summarized in the *common disease-common variant* hypothesis, which argues that common, rather than rare, genetic variants underlay the heritability of common diseases. If this hypothesis is true then it has been shown [14] that genetic association studies will have more power than

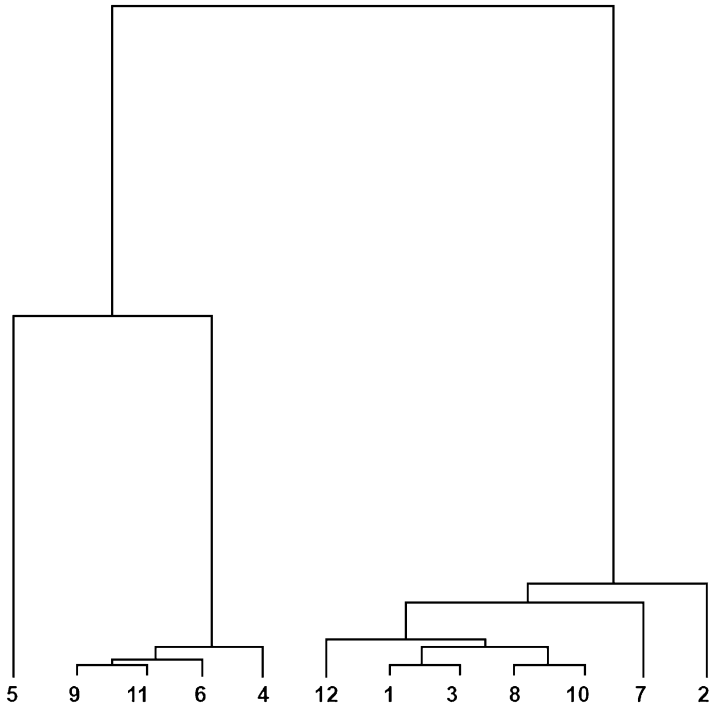

Fig. 2.1 Simulation of SNP data through the coalescent process

traditional pedigree-based linkage studies when it comes to finding the risk alleles. The arguments for the common disease-common variant hypothesis (c.f. [15–18]) include (1) weak or nonexistent selective pressure against variants involved in susceptibility for common diseases (especially late onset diseases which would tend not to affect reproductive success), (2) expansion of populations after passing through *bottlenecks* related to migration away from the earliest human groupings, and (3) a resulting distribution of alleles that involve susceptibility in which common variants predominate. The latter point follows from the first two; the distribution of frequencies of selectively neutral alleles in stable populations was described by Wright [19], further refined by Ewens [20], and has been updated to account for population growth and selection [17, 21].

An important implication of the neutral selection model is that, while there are many rare markers, if two individual chromosomes are compared to one another, most of the differences seen between the two chromosomes are due to common, rather than rare, alleles. Specifically, considering a specific site where the two genotypes differ, if we take a larger sample of individuals and examine the same site in each of the new individuals, in most cases many of them too will also be found to differ at this site as well. A heuristic argument for this can be made by considering the time to coalescence between the two individual chromosomes showing a different allele at a specific site. In order for the two sequences to differ

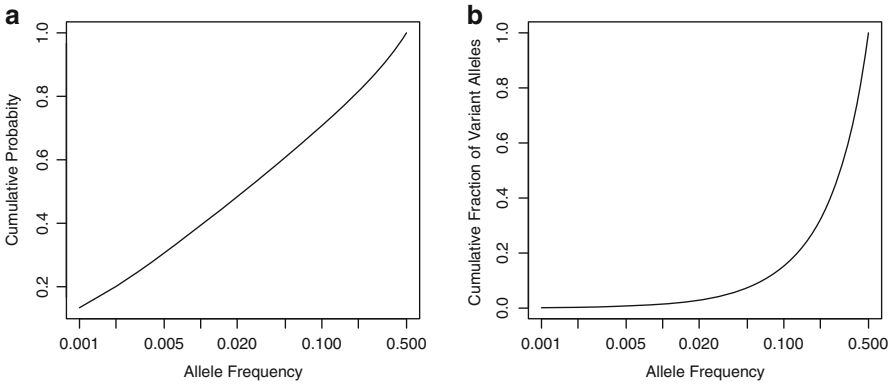

Fig. 2.2 (**a**) Expected cumulative distribution of allele frequencies of variants seen when genotyping a large number of sites in a sample population of size 500. (**b**) Expected cumulative distribution of total minor allele counts in same study

at a given site, the mutation at that site must have occurred after the time of coalescence for the two chromosomes (here we are focusing on segments short enough not to have been affected by recombination). The mean coalescent time is N generations in the past with an expected time of appearance of the mutation equal to $1/2N$ (assuming a constant rate of mutation over time). But this is early enough so that it would predate a very large number of coalescent events in the larger sample since the expected time to the MRCA for all the subjects in the sample is $2N(1 - 1/n)$ or roughly $2N$ for large n; in fact on average we would expect, from (2.11) and (2.12), that the last 3 coalescences take up time $3/2N$ so that all but these are (expected to be) yet to come by the (expected) time of the observed mutation. Thus a large fraction of the n members of the sample are likely to carry the mutation of interest. There is considerable variability in both the actual times of mutation and the number of coalescent events yet to take place at that time, but this argument can be fleshed out in further detail using coalescent theory.

For example, a formula for the expected number of sites, η_i, at which the less frequent base is present on i sequences out of a sample size of n sequences ($N = n/2$ individuals) is given by Wakeley [3] [(4.21), p. 105)] as

$$E(\eta_i) = \theta \frac{\frac{1}{i} + \frac{1}{n-i}}{1 + \delta_{i,n-1}} \quad 1 \le i \le [n/2], \tag{2.13}$$

where θ is the (scaled) mutation frequency, $[\cdot]$ is the largest integer (i.e., floor) function, and δ_{ij} is the Kronecker δ function equal to 1 if i and j are equal and equal to 0 otherwise. Figure 2.2a shows the expected cumulative frequency of minor allele frequencies for SNPs seen at least once in a sample of 1,000 chromosomes (500 diploid individuals) using this formula (by considering only SNPs seen at least once, we can drop the mutation frequency parameter from the calculations).

It is clear from Fig 2.2a that a large proportion of variants are expected to be "rare," with, for example, about 39 % of total alleles having allele frequency less than one percent while only about 20 % of variants will have frequency between 0.2 and 0.5. However, things look quite different when counting the total number of variant (e.g., minor) alleles seen, since each variant contributes an expected $2pN$ minor alleles; Fig 2.2b gives the expected cumulative contribution to the total number of minor alleles seen according to allele frequency. SNPs with frequency less than 1 % contribute hardly at all to the total variation, while common SNPs (frequency 0.2–0.5) contribute approximately two thirds.

This same argument applies to genetic variants (causal alleles) that increase susceptibility to diseases that have no relationship to reproductive success, either because they do not interfere with fitness or because they tend to be too late in onset to affect fertility. As will be shown in Chap. 7, common causal alleles are inherently more detectable than are rare alleles, all other things being equal, and would seem to contribute far more to risk of disease than rarer SNPs, unless there is a marked tendency for rarer variants to have higher risks associated with them.

2.6.4 Allele Distributions Under Population Increase and Selection

One factor in human population history that influences the number of rare variants (and potentially their contribution to risk) relative to the discussions above is the very large growth in human populations in the last 2–3 hundred years. This recent spectacular growth implies an excessive fraction of alleles are rare compared to the model given above. Recent large-scale sequencing in the *1000 Genomes Project* show that there is a surfeit of SNP alleles with frequency less than 0.5 % compared to that expected under the constant population size model [22].

An additional, and potentially more important factor influencing allele frequencies, is selection. Equation (2.13) is based on an assumption that all the alleles considered are neutral in the effects on reproductive success. A formula for allele frequency distributions in an infinite population that includes the effect of natural selection was given by Wright [23] who found that allele frequency p is distributed as

$$f(p) = kp^{(\beta_S - 1)}(1-p)^{(\beta_N - 1)}\mathrm{e}^{\sigma(1-p)}, \tag{2.14}$$

with β_S as scaled mutation frequency (equivalent to θ above), β_N as scaled back-mutation (reversion) frequency, σ as the scaled selection rate, and k as a normalizing constant. Note that $f(p)$ in (2.14) is not really a probability distribution since its integral does not exist over the range (0–1) of interest. However, we can still use Wright's formula to evaluate questions involving the relative frequencies of ranges of alleles in an infinite population. The key parameter of interest to us here is the selection parameter σ; this is given in units of $4N_e s$ with N_e the effective sample size and s the reduction in reproductive success (expressed as the fractional reduction in

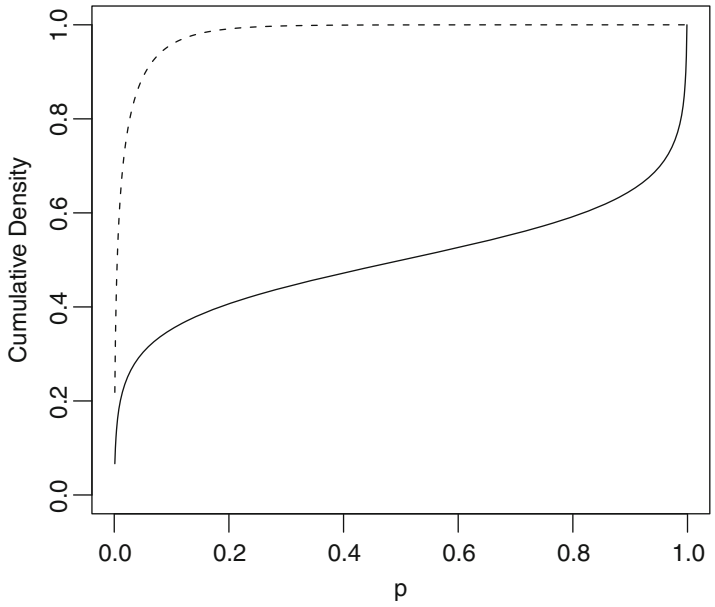

Fig. 2.3 Allele frequency distribution under selection or not. Wright's formula is used to depict the cumulative density of allele frequencies between frequency 0.001 and 0.999 under no selection (*solid line*) and selection with n_b (*dotted line*)

the number of offspring expected from carriers compared to noncarriers) associated with the allele of interest. Following Pritchard [24], we set σ to equal 12 to signify "weak selection" against the allele of interest and zero for no selection. Note that $\sigma = 12$ translates to a 0.03 % reduction in the probability of leaving offspring in each generation assuming a constant effective sample size of $N_e = 10{,}000$. Figure 2.3 plots the cumulative distribution of allele frequencies under this model. Even under this weak level of selection pressure, it is evident that there is a dramatic shift to rarer alleles compared to the neutral model. Does this mean that the *common disease-common variant* hypothesis that has underpinned GWAS studies to date is misconceived? This and related questions is taken up in Chap. 8.

2.7 Recombination and Linkage Disequilibrium

Note that in the coalescent depicted in Fig. 2.1 if two mutations (call them A and B) occur along a specific path segment between coalescences then in the resulting population all chromosomes that carry mutation A would also carry mutation B as well. This observation illustrates the first of two concepts (linkage disequilibrium) underlying the association-based approach to identifying risk alleles. The second is recombination.

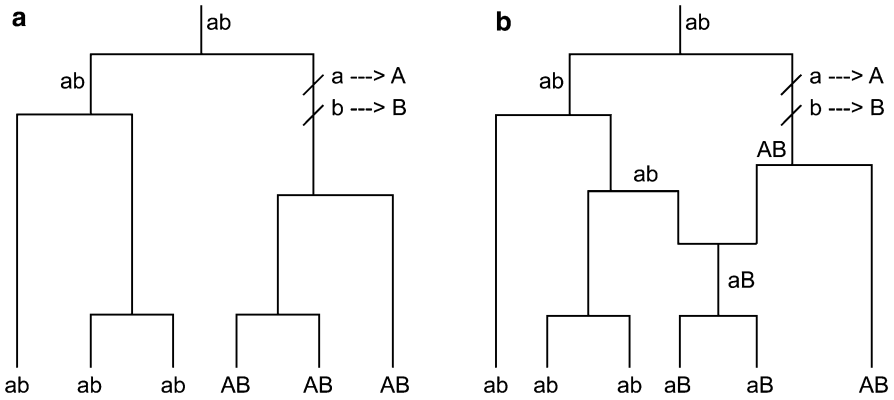

Fig. 2.4 Recombinations and the ancestral recombination graph. Two mutations are depicted as having occurred on the same branch. (**a**) No recombination. (**b**) Recombination that breaks the correlation between mutations A and B

Recombination is the exchange of genetic material during meiosis as two (homologous) chromosomes split at a certain point and reorganize after having traded the DNA sequence on one or the other side of the split with each other. Now markers *A* and *B* are no longer traveling together in the progeny issuing from that point forward thus altering (weakening) the pattern of correlation between these two markers. Approaching this backwards in time we note that when a recombination occurs, we can think of the child chromosome as having chosen two parents at that generation time, *t*, i.e., one for the position that *A* occupies and one for the position that *B* occupies. Figure 2.4 gives an illustration of such an ancestral recombination graph [25]. The ancestry of each position (one for *A* and one for *B*) marginally follows a coalescent. Ultimately even with recombination, there will be a single MCRA found from which both loci were inherited [25] and ultimately a single MCRA for all sites for the entire chromosome and genome, i.e., common ancestors for all living humans.

If the recombination occurs early in the coalescent process, the correlation between *A* and *B* will be weakened possibly considerably since a large number of both recombinant and nonrecombinant chromosomes may have progeny in the current generation. On the other hand, if the recombination event occurs much later down the tree than the initial occurrence of these two mutations, most, but not all, of the resulting (modern day) chromosomes that carry mutation *A* will also carry mutation *B*, i.e., the correlation between *A* and *B* will have been only partly diminished by the recombination. Here we have considered the impact of just one recombination event on the resulting correlations, but in fact if *A* and *B* are widely separated then many recombination events will generally have occurred, each one will tend to reduce the association between alleles like *A* and *B*.

The basic principles underlying association-based genetic studies is that markers are correlated with each other and with causal variants of any type which have

similar histories, for example, the allele A may be a variant that increases susceptibility to disease and B is a marker on the same chromosome that originated on the same path segment (as in our simplified example) in Fig. 2.1. If there were no recombination events anywhere on an entire chromosome, then marker B would not be informative at all about the location of risk allele A; i.e., B could be placed anywhere on the same chromosome and still be highly correlated with A, so that B would appear to be associated with disease. However, when recombination does occur, the correlations between markers are weakened. If the loci occupied by A and B are distant from each other, then there will have been many recombinations taking place over time so that the correlation between the marker B and the risk allele A will be so low that B will no longer appear to be related to the disease of interest.[2]

The probability, in any generation, of a recombination occurring between the loci that are occupied by A and B is termed the *genetic distance* between these two sites. Genetic distance is an increasing function of the distance between sites, but not uniformly so because recombination events do not occur at uniform rates at all positions on a chromosome. There are certain regions where recombinations are rare and certain regions where they are unusually common. These may be termed recombination cold spots or hot spots, respectively [26–29]. Alleles between which there have been low numbers of historical recombinations are said to be in linkage disequilibrium (LD). The ongoing revolution in biotechnology has allowed the inexpensive genotyping of hundreds of thousands of markers in large numbers of people; the number of markers is so large that there is likely to be at least one marker in high linkage disequilibrium with any given (common) variant affecting risk of a disease or influencing some other phenotype that we are interested in. Depending on the frequency of both the marker and the underlying causal variant, this may result in a statistically detectable association between the disease of interest and the marker. Once such an association has been detected, the chromosomal region surrounding the marker becomes a candidate for a number of additional analyses designed to first replicate the association, then further localize and identify variants, and ultimately characterize the mechanism by which a causal variant affects the disease or phenotype.

2.7.1 *Quantification of Recombination*

There are a number of standard statistics that are used to describe the pattern of association between markers and to therefore characterize the extent of linkage disequilibrium between markers. The two most important of these are Lewontin's

[2] Note that we are now talking about correlations between SNP allele counts of two different markers at different locations, over a sample of individuals, $i = 1,\ldots,N$. In Sects. 2.1–2.5, our focus was on the correlations (induced by structure or relatedness) of counts, n_A of a single marker A for different individuals i and j.

Table 2.4 Haplotype probabilities for two loci

	b	B	
a	$(1-p)(1-q)+\delta$	$(1-p)q-\delta$	$1-p$
A	$p(1-q)-\delta$	$pq+\delta$	p
	$1-q$	q	1

[30] δ', while the second is the usual R^2 criterion applied to alleles at two loci. These two statistics are closely related; the first is designed to be a measure of linkage disequilibrium between the two independent of their allele frequencies, while the second depends on both the degree of linkage disequilibrium and the allele frequencies of the two markers. The R^2 statistic is most directly related to the performance of association-based genetic studies in the situation we described above, namely, when a nearby marker is measured as a surrogate for an unmeasured (and generally unknown) causal variant. Both δ' and R^2 can be initially considered in terms of a 2×2 table of two diallelic variants at two loci on a single chromosome. Each of the four possible combinations of the two alleles for the two variants is termed a *haplotype*. If we denote the two variants as having alleles a or A for the first variant and b or B for the second and if p is the probability of allele A (so that $1-p$ is the probability of allele a) and q the probability of allele B, then the probabilities of the four haplotypes, ab, aB, Ab, and AB, can be depicted as in Table 2.4.

Here δ is a disequilibrium parameter that measures the association between the two loci in the sense that if δ is zero, the two alleles are independent and otherwise are positively (for $\delta > 0$) or negatively ($\delta > 0$) associated. In fact, we see below that δ is the covariance in the population of haplotypes of the counts (0 or 1) of the number of alleles at each of the two loci. Because all haplotype probabilities must be nonnegative, we can immediately see that δ is bounded by $-\min((1-p)(1-q), pq)) \le \delta \le \min(p(1-q), (1-p)q)$. Since the labeling of the alleles (a vs. A or b vs. B) is completely arbitrary, yet affects the sign of δ, the interest is actually in $|\delta|$ which is bounded by

$$\delta_{\max} = \begin{cases} \min\left[q(1-p), (1-q)p\right] & \text{if } \delta > 0 \\ \min\left[(1-p)(1-q), pq\right] & \text{if } \delta < 0 \end{cases}.$$

The scaled version of $|\delta|$ is defined as $\delta' = (|\delta|/\delta_{\max})$ which takes values between zero and 1. Note that if δ' takes its maximum value of 1, then this means that at least one cell (i.e., haplotype) in Table 2.4 will have probability zero. Note that if all haplotypes have nonzero probability, then this constitutes proof (under the infinite sites model) that there has been a historical recombination between the two loci, i.e., the only way that all four haplotypes, ab, aB, Ab, and AB could occur is if there was either (1) a second de novo occurrence of either A or B (assuming that ab is the ancestral allele), ruled out by the infinite sites model, or (2) that there was a recombination between the two loci to form, for example, AB from aB and Ab.

While a value of δ' close to one is evidence that there has been little or no recombination between the two loci δ' does not really measure the strength of the association between the two variant alleles in a way that is important for association testing. For example, if the allele frequencies p and q are very different (e.g., if p is close to 1/2 but q is much smaller), then the presence of only three haplotypes, for example, ab, aB, and Ab, means that while A always appears with b and never B, there are many instances of b not associated with A. When it comes to assessing the power of an association study that genotypes the loci occupied by B but not the loci occupied by A, the appropriate linkage disequilibrium criteria is simply the squared correlation R^2 between the count of the two alleles A and B.

Working within the haplotype framework, we introduce a slight variation on previous notation so that $n_A(h)$ counts the number of A alleles on haplotype h. Thus for haplotypes ab, aB, Ab, and AB, $n_A(h)$ takes values 0, 0, 1, and 1, respectively; similarly $n_B(h)$ counts the B allele.

From first principles $R^2 = \text{Cov}[n_A(h), n_B(h)]^2/\{\text{Var}[n_A(h)]\text{Var}[n_B(h)]\}$. The variance of the count, n_A, of the number of alleles of A (now a binary random variable since we are dealing with only one haplotype) is simply $p(1 - p)$ and for n_B is $q(1 - q)$, with the covariance between the two equal to $E[n_A(h)n_B(h))] - E[n_A(h)]E[n_B(h)] = (pq + \delta) - pq = \delta$ so that $R^2 = \frac{\delta^2}{p(1-p)q(1-q)}$. It is worth noting first that R^2 does not depend upon which of the two alleles, a or A, is being counted, i.e., we could use n_A or n_a equivalently, and similarly for n_B or n_b.

2.7.2 Phased Versus Unphased Data and LD Estimation

While we have described R^2 as the squared correlation between binary variables (the n counters) in an experiment where we observe the haplotypes, ab, Ab, aB, and AB, directly, the same R^2 (or R for that matter) applies if we observe, rather than the haplotypes themselves, unphased genotypes for diploid chromosomes, so long as we assume that each haplotype is inherited independently; in this case n_A and n_B are correlated binomial random variables, which now take the values (0, 1, or 2), but with the same value of the correlation between them as if haplotypes were measured directly. Since we can certainly estimate the allele frequencies p and q as well from the unphased as from the haplotype data, a non-iterative estimate of δ based solely on the genotype data is simply $\hat{\delta} = \hat{R}\sqrt{\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})}$. Here we have estimated R as the usual sample correlation estimate for n_A and n_B from genotypes obtained for a sample of individuals in the population. Other estimates of δ can be formed from genotype data, with the method of maximum likelihood easily implemented using an EM algorithm [31]. More about estimation of haplotypes based on genotype data is presented in Chaps. 5 and 6.

2.7.3 *Hidden Population Structure*

As we will see in Chap. 3, genetic association studies are based upon the assumption that because of recombination only markers that are in close proximity to a causal genetic variant that directly or indirectly influences disease risk or phenotype distribution should be statistically associated with that phenotype. However, there are a number of reasons why this may not be true. The most important of these is hidden population structure present in the study. When hidden population structure is present, LD appears to be of much greater extent than expected, and very large numbers of markers scattered over the entire genome may show unexpected levels of association with the phenotype of interest. At the root of this failure to localize associations is the extent of apparent LD that is brought about by the presence of such phenomenon as admixture, hidden stratification, and cryptic relatedness.

2.7.3.1 Stable Populations

Consider first a stable population of chromosomes that is isolated, in that there is no migration from the outside entering the region, and where there is random mixing between chromosomes (i.e., genetic variants are all assumed to be neutral in terms of reproductive fitness and unrelated to mating choice). In this case over time we will see two things: (1) an increase in the number of (neutral) mutations and (2) a loss of linkage disequilibrium between existing markers. For a given pair of markers t, it can be shown that the linkage disequilibrium parameter δ changes with generation, t, according to

$$\delta(t) = \delta_0(1-\theta)^t, \tag{2.15}$$

where δ_0 is the initial value of δ and θ is the probability of a recombination between the two markers in one generation. Ultimately with enough time elapsed, recombination between markers would insure that only those markers that are extremely close to each other physically (i.e., with θ very close to zero) should be in LD with each other. As we will further illustrate later in this chapter, of the major continental populations (e.g., European, African, and Asian), it is the oldest continental population, i.e., people in or from Africa, that tends to show the lowest LD between markers and also the most common variants.

2.7.3.2 Out Migration and Population Expansion

Groups of Africans began to expand into Europe and Asia at the end of the last great ice age. Outmigration of a small number of founders and subsequent rapid population growth has profound effects upon linkage disequilibrium and frequency of markers. Since the founders may each have a great number of descendents, any mutation carried by a founder, no matter how rare in the ancestral population, will

tend to be common in the descendant population. Moreover mutations that were in linkage equilibrium in the ancestral population, but which were both carried by a given founder chromosome, will be in linkage disequilibrium in the new population and remain in linkage disequilibrium until the force of recombination again weakens such disequilibrium. Thus recently founded isolated populations will tend to have higher levels of linkage disequilibrium than will older more stable populations.

For GWAS studies the extended linkage disequilibrium of recently derived isolated populations has sometimes been considered to be of benefit in that it would take fewer markers to find regions of the genome that were associated with disease- or phenotype-associated causal variants. Iceland, for example, has been described [32, 33] as an ideal setting for conducting GWAS studies because in this recently derived ethnically homogeneous island, LD patterns should be longer and coverage of causal variants with GWAS chips should therefore be more complete. In fact the GWAS studies in Iceland, conducted by deCode Genetics corporation, has had many important successes. It appears however that these successes had much more to do with the fact that DNA resources for a large fraction of the entire population of Iceland were available at an early stage, as well as the abilities of the deCode scientists, and less to do with better coverage of the genome, at least in comparison with other European populations. In particular, the discoveries in Iceland were very often reproduced by other similar sized GWAS studies of other less recent but still European-derived populations using similar genotyping technology.

A downside of using an isolated population in a GWAS is that not all alleles related to the phenotype of interest disease are likely to be present in that population (due to founder effects), and moreover extended linkage disequilibrium implies that localization of causal alleles underlying observed associations may be more difficult. Studies of disease and other phenotypes in populations of more ancient origins, notably African-derived groups, do require more SNPs to cover any one particular genetic region, or the genome as a whole, and this was one early impediment to progress in GWAS studies in such populations. The ability of commercial GWAS platforms to cover the variation in the African genome is gradually improving, and the shorter LD can be beneficial for the localization of the signal from GWAS findings, a crucial step in the ultimate identification of specific causal variants. The most crucial impediment now to GWAS studies in African populations is the less developed infrastructure of cohort and case–control studies with access to DNA samples in this (and other) non-European groups.

2.7.3.3 Population Admixture and Hidden Stratification

In the last few hundred years not only have population sizes grown immensely but also groups that had been relatively isolated for long periods are now in contact, because of migration patterns brought on by technological advances, the growth in trade and other economic relations between groups, and for less positive reasons including war, conquest, and slavery. Recent admixture between formerly isolated

groups is an important feature of many of today's modern populations. Recent admixture introduces linkage disequilibrium between markers at two loci which have different allele frequencies in the original populations; this linkage disequilibrium is extensive and even affects markers that are on different chromosomes; if the resulting admixed population then mixes freely in subsequent generations, truly long-range linkage disequilibrium (between markers separated so that their recombination rate is near 1/2) will disappear rapidly , as indicated by formula (2.15), in subsequent generations. If mating practice is affected by admixture proportion (with mating more likely between similarly admixed than dissimilar subjects), then apparent linkage disequilibrium can remain high between markers throughout the genome for many generations. Admixture is an important aspect of such US populations as African Americans, Latinos, American Indians, and Native Hawaiians, and additional admixture among many populations may be expected to continue due to continued population movements and loosening of cultural prohibitions. As we will see in Chap. 4, it may be very important to control for admixture in an association study.

Hidden stratification occurs when an apparently homogeneous population contains unrecognized (or simply unmeasured in a given study) population structure, specifically the presence of subgroups within a larger population between which mixing is rare. Historically within the USA, we can think of many religious or ethnically based groups with prohibitions against marriage outside the group, but which may not be distinguished from each other in typical epidemiological studies. The distinction between hidden population stratification and admixture is not very clear in practice since in many cases traditional prohibitions are increasingly relaxed. The impact on LD structure of such stratification (while usually milder than the effect of recent admixture between continentally separated groups) can still be important in association studies especially in certain types of extreme studies, as we will see in later chapters.

2.7.3.4 Hidden Relatedness Between Subjects

Another potentially problematic issue for association studies is the presence of unexpected close relatives in studies. Many association studies that enroll a large number of participants from small populations are likely to encounter a fairly large number of quite closely related subjects, and even large-scale multisite studies occasionally find unexpected relatives (and even identical twins) when they look carefully at the genotypes that are generated. As we see in Chap. 4, *cryptic relatedness*, does, if many close relatives are genotyped in a study but the subjects were not recognized to be close relatives, have an important effect on the distribution of test statistics designed to detect associations between markers and phenotype.

While relatedness between individuals does not affect the marginal distribution of individual genotypes (here we are ignoring the possibility of inbreeding), it does affect the distribution of quantities such as the sum of two or more related individual's genotypes, for example, for two siblings (so that $z_0 = \frac{1}{4}$, $z_1 = \frac{1}{2}$, and $z_2 = \frac{1}{4}$) so that

the correlation between n_1 and n_2 is equal to 1/2 $z_1 + z_2 = 1/2$, then the sum, S, of $n_1 + n_2$ will have variance

$$\mathrm{Var}(n_1) + \mathrm{Var}(n_2) + 2\mathrm{Cov}(n_1, n_2) = 6p(1-p)$$

compared to $4p(1-p)$ if the two counts were independent. This comparison is important because, as discussed in Chaps. 3 and 4, statistical tests that use inappropriate variance estimates are generally overdispersed under the null hypothesis, i.e., are prone to give inappropriate false-positive (type I) error rates.

2.7.4 Pseudo-LD Induced by Hidden Structure and Relatedness

For most of this chapter only one marker has been considered in the discussion of hidden structure and relatedness. Now consider an extension of the Balding-Nichols model to two markers, not in linkage disequilibrium in any individual population, sampled for a set of individuals ($i = 1, \ldots, N$) in a stratified population with substrata $l = 1, \ldots, L$. For the purpose of notational simplification, assume that both markers have the same frequency, p, in the ancestral population. The simulation process is as follows:

For the first marker, draw a set of subpopulation marker frequencies, p_{l1}, $l = 1, \ldots, L$ from the beta distribution with mean p.

For the first marker, draw a sample of marker values $n_{i,1}$, $i = 1, \ldots, N$ independently from the binomial distribution with index $= 2$ and frequency p_{l1}, where l corresponds to the substratum that individual i is assigned to.

Draw a set of marker frequencies, p_{2l}, $l = 1, \ldots, L$ from the same beta distribution for the second marker.

Draw sample of the second marker $n_{i,2}$, $i = 1, \ldots, N$ again stratified by subpopulation.

The following *R* code snippet illustrates the sampling procedure for $L = 2$ populations each of size 1,000 subjects:

```
> set.seed(10201) # to get the same results each time
> N<-1000 # number of subjects in each of L strata
> L<-2    # number of strata
> p=.3    # ancestral allele frequency for both markers
> F=.1    # differentiation between modern-day and ancestral populations
> n1<-rep(0,L*N) # holds values for first marker
> n2<-rep(0,L*N) # holds values for second marker
> p_l1=rbeta(L,p*(1-F)/F,(1-p)*(1-F)/F) # freqs of marker 1 in current populations
> p_l2=rbeta(L,p*(1-F)/F,(1-p)*(1-F)/F) # freqs of marker 2 in current populations
> for (l in 1:L){ # sample from appropriate binomial distribution
>      n1[((l-1)*N+1):(l*N)]<-rbinom(N,2,p_l1[l])
>      n2[((l-1)*N+1):(l*N)]<-rbinom(N,2,p_l2[l])
> }
```

Now consider testing for whether there is an association between markers 1 and 2. After running the above code, we type:

```
> summary(lm(n1~n2))
```

which gives results:

```
Call:
lm(formula = n1 ~ n2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.69113    0.02006  34.456  < 2e-16 ***
n2          -0.06808    0.02122  -3.208  0.00136 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that variables n_2 and n_1 appear to be associated despite being sampled independently, while the magnitude of the correlation is not very large (just $R = -0.07$), the p-value for testing association is equal to 0.00136 clearly something that is unlikely just by chance, in fact repeating this experiment many times gives a large number of quite small p-values.

The real source of this apparent association between independent markers is the failure to take account of the correlation structure of n_1 when we tested for association between n_1 and n_2.

The *R* function *lm* has underestimated the variance of the ordinary least squares estimate because it has assumed that the elements of n_1 are independent of each other, when instead the nonzero value of F used in the simulation has induced the between-marker correlation structure described above. When averaging over many runs, the correlation will be estimated to be equal to zero between the two markers; however, for any given run the usual tests for association are very likely to yield a very small p-value, i.e., a false-positive assessment of correlation, and this problem increases with sample size N.

A more detailed analysis (see Chap. 4) shows that the assumption that both n_1 and n_2 are differentiated according to population structure is required to produce this sort of induced correlation (this makes sense since we can always reverse the role of n_1 and n_2 in this problem). Note the use here of ordinary least squares (the *R* procedure *lm*) rather than *glm* to test for a relationship between two marker count variables; while this may seem nonstandard to many statisticians familiar with analysis of binomial data (e.g., using logistic regression), it is quick and gives very well-behaved tests of the null hypothesis of no association between the two markers for sample sizes as large (1,000 per strata) as used here (this can be observed in the behavior of the simulation program by setting F to zero and running the entire experiment again—only something very close to the expected number of small p-values will be observed), see Chap. 3 for more information.

Clearly the induced correlation described here behaves nothing like true linkage disequilibrium, which should die out as the genetic distance between the two markers increases; here we have simulated a situation that is analogous to markers lying on different chromosomes entirely. Interestingly as we will see in Chap. 4,

even such weak apparent correlations (because they are spread over the entire genome) can have important consequences when it comes to distinguishing true marker/genotype associations between those induced by hidden structure or relatedness. This is most especially true for certain extreme analyses such as the case-only method for assessing gene x gene interactions [34] (see Chap. 7).

2.8 Covering the Genome for Common Alleles

As described above the common disease-common variant hypothesis is undoubtedly a simplification of a much more complex genetic architecture for many diseases, and while common disease-related alleles have been detected for many diseases, many questions remain about the relative role of common variants versus rare variants. Nevertheless the CDCV hypothesis motivates the use, in genetic association studies, of the technical breakthroughs that have led to the identification of millions of common SNPs, as well as cost-effective array-based large-scale genotyping of hundreds of thousands or even millions of SNPs.

A key issue then for GWAS studies is the "coverage" of common variants by the SNP arrays that are available commercially; coverage calculations are based on R^2 statistics between SNPs on the arrays and a target set of known variants. For each target variant, the goal is to have an SNP on the SNP array that is highly correlated with the target. Since these arrays are meant to be used generally, i.e., for all heritable phenotypes that could be of interest, target variants should not be simply be SNPs in candidate genes for one or more phenotypes but rather should as much as possible constitute a census of all common variation.

SNP array coverage statistics have been described for various populations by the companies that manufacture and sell the arrays, as well as by other investigators [35, 36]. However, at least three factors make the evaluation of coverage of the genome an ongoing project; first of all, the frequency domain of the target variants (which originally specified variants with minor allele frequencies at least 5 to 10 % frequency in one or more populations as common) is changing with rarer alleles increasingly being considered to be worthy targets; secondly the number of populations being assessed for genetic variation is increasing and this can lead to additional variants, common in specific populations being included in the target set; and finally techniques of SNP discovery are continuing to improve and to lead to increased discovery of variants not known to have existed. One additional factor that may also complicate the assessment of the coverage of common variation is inherent uncertainty in the measurement of some kinds of potential causal variants. Here I am referring to the possibility that haplotypes made up of SNPs or other variants could be causal variants (examples have been reported, e.g., [37]) rather than the SNPs: a "haplotype effect" means that a combination of SNPs falling on the same homologous chromosome could have a different effect than if the same SNPs were present, but on different

homologs. Even if all the SNPs that constitute a SNP haplotype appear on the array, there still may be (depending on the degree of historical recombination between the SNPs making up that haplotype) considerable uncertainty in inferring the haplotype based upon the genotypes of those SNPs. We discuss haplotype uncertainty in more detail in Chap. 7. If we begin to consider not only all common SNPs and other sequence variants but also *all common haplotypes* of these SNPs as our target set of variants, then this greatly increases both the number of variants to be tested and also lowers overall coverage statistics due to haplotype uncertainty.

To date, the coverage target set of choice has been the SNPs (and other identified inherited variants) genotyped in the HapMap project (http://hapmap.ncbi.nlm.nih.gov/), specifically in phase 1 and phase 2 of this project [38]. As described below, phase 1 and phase 2 together genotyped over six million SNPs of which several millions were found to be common in at least one of four groups of samples. The DNA samples genotyped by the HapMap project originated from 270 people: 90 Africans (members of the Yoruba people, in Ibadan, Nigeria), 90 Americans of European descent from Utah, 45 Han Chinese from Beijing China, and 45 Japanese from Tokyo, Japan. The SNPs that were chosen to be genotyped in HapMap were those that had been previously reported to dbSNP (http://www.ncbi.nlm.nih.gov/snp/ [39]), an online repository of information about genetic variation contributed by scientists studying the human genome from around the world. In order to assess completeness of ascertainment of common SNPs by the HapMap, the HapMap project included within it an SNP discovery project focusing upon 10 different regions of the human genome, each of approximate length 500 kb, for which genetic sequencing was performed in 48 unrelated DNA samples (16 Yoruba, 8 Japanese, 8 Han Chinese, and 16 Europeans); all identified SNPs as well as all SNPs from dbGAP reported in these regions were genotyped in all the HapMap samples. Reports about the frequency and number of common variants in these regions as well as other statistics designed to use these regions to assess coverage over the full genome have been published [38].

More recently the 1000 Genomes Project [40] is in the process of using high-throughput sequencing methods to identify additional common sequence variants and also to begin to for the first time create a census of less common variants, specifically those within the range from 1 to 5 % frequency in at least one of the 14 populations considered. While these data are still (at the time of this writing) incomplete, there are already indications that not all common variants to be found in this project will have good surrogate SNPs either on the commercial arrays or indeed within the HapMap itself.

Table 2.5 shows the number of individual samples and type of sample (unrelated individuals or parent-offspring trios) for the HapMap phase 1 + 2 project, the HapMap phase 3 project, and the 1000 Genomes Project.

Table 2.5 Population description for HapMap and the 1000 Genomes Project (1KP)

Population designator	Population descriptor	HapMap 1 + 2	HapMap 3	1KG phase 1
ASW	African ancestry in southwest USA			
	N samples	0	90	61
	N founders	0	60	61
CEU	Utah residents with northern and western European ancestry			
	N samples	90	180	85
	N founders	60	121	85
CHB	Han Chinese in Beijing, China			
	N samples	45	90	97
	N founders	45	90	97
CHD	Chinese in Denver, Colorado			
	N samples	0	100	0
	N founders	0	100	0
CHS	Han Chinese, South			
	N samples	0	0	100
	N founders			100
CLM	Columbians in Medellin, Columbia			
	N samples	0	0	60
	N founders			60
FIN	Finnish in Finland			
	N samples	0	0	93
	N founders			93
GBR	British from England and Scotland			
	N samples	0	0	89
	N founders			89
GIH	Gujarati Indians in Houston, Texas			
	N samples	0	100	0
	N founders	0	100	0
IBS	Iberian populations in Spain			
	N samples	0	0	14
	N founders	0	0	14
JPT	Japanese in Tokyo, Japan			
	N samples	45	91	89
	N founders	45	91	89
LWK	Luhya in Webuye, Kenya			
	N samples	0	100	97
	N founders	0	100	97
MEX/MXL	Mexican ancestry in Los Angeles, CA			
	N samples	0	90	66
	N founders	0	60	66
MKK	Maasai in Kinyawa, Kenya			
	N samples	0	180	0
	N founders	0	150	0

(continued)

Table 2.5 (continued)

Population designator	Population descriptor	HapMap 1 + 2	HapMap 3	1KG phase 1
PUR	Puerto Ricans in Puerto Rico			
	N samples	0	0	55
	N founders	0	0	55
TSI	Tuscany in Italia			
	N samples	0	100	98
	N founders	0	100	98
YRI	Yoruba in Ibadan, Nigeria			
	N samples	90	180	88
	N founders	60	120	88

2.8.1 High-Throughput Sequencing

As of the time that this book is written, large-scale high-throughput sequencing [41] is beginning to be performed in association studies either focused on candidate regions, such as exons either of candidate genes, or for all known genes, regions containing GWAS hits; whole genome sequencing is beginning to be used despite the considerable expense and informatic burdens that large-scale sequencing involves. More on the rationale for collecting and approaches to the analysis of such data is given in Chap. 8.

2.9 Principal Components Analysis

One of the most useful methods for detecting and visualizing hidden structure and admixture is principal components analysis [42], and this technique has been widely advocated (and used) as an effective tool for addressing population structure in the analysis of data from association studies [4]. For the time being, we concentrate on principal components methods as a technique for display of sample or population features.

Reviewing very briefly, see Jolliffe [43], given samples of random vectors X with M components, the first principal component of X is a linear function $\alpha'_1 X$ of the elements of X which has maximum variance under the constraint that $\alpha'_1 \alpha_1$ equals 1. The second principal component is another linear function, $\alpha'_2 X$, which is uncorrelated with $\alpha'_1 X$ and again with maximum variance subject to $\alpha'_2 \alpha_2 = 1$. Proceeding in this fashion further up to M principal components can be found, but it is hoped that important characteristics of the variability of X will be captured by L principal components with L considerably less than M. These first L linear functions are called the *leading* principal components. They can be computed through a *spectral* or *eigenvector*/*eigenvalue* decomposition of the covariance

matrix of X; specifically if $\mathbf{\Sigma}$ is the covariance matrix of X, then the *spectral decomposition* of $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \sum_{k=1}^{M} w_k \alpha_k \alpha_k^{\mathrm{T}},$$

where the w_k are each nonnegative scalars (the eigenvalues of $\mathbf{\Sigma}$), with $w_1 \geq w_2 \geq w_3 \ldots \geq w_M \geq 0$, and the α_k as the eigenvectors of $\mathbf{\Sigma}$ each having the property that $(\mathbf{\Sigma} - w_k \mathbf{I})\alpha_k = 0$. The fraction of total variance that is explained by the L principal components is computed as

$$\sum_{i=1}^{L} w_i \Big/ \left(\sum_{i=1}^{M} w_i \right). \tag{2.16}$$

Principal components analysis is used as a data summarization technique in a wide variety of data intensive fields and disciplines including bioinformatics for gene expression data, proteomics, and metabolomics [44, 45]. In nutritional epidemiology, there have been a number of papers relating risk of late onset disease (cardiovascular disease, cancer, etc.) to specific eating patterns captured by the first L (or leading) principal components of dietary data [46]. Many other uses of principal components analysis and its cousin, factor analysis, have been described in many different fields including bioinformatics, meteorology [47], and economics.

When variables are not all scaled similarly, it is customary to compute the principal components from the correlation matrix for the random vector X rather than from the covariance matrix. When dealing with massive amounts of SNP data (i.e., where M corresponding to the number of SNPs genotyped in a given study is very large), large-scale genetic association studies tend to approach principal components slightly differently than described above. For example, Price et al. [4] describes the extraction of the eigenvectors of the $N \times N$ kinship matrix $\mathbf{K}$, where N is the number of subjects, rather than starting with (as above) the eigenvectors of the $M \times M$ correlation or covariance matrix for the SNP data. The first L of these eigenvectors themselves (not their dot products with each individual's SNP data) are then termed "principal components" by Price et al., i.e., the ith element of the jth eigenvector is referred to as the value of the jth "principal component" for person i. While it is nonstandard terminology to call these the principal components, there is actually a close relationship between the classical principal components, calculated by decomposing the correlation matrix of the genotype data, to the eigenvectors of the relationship matrix (a matrix of dimension $N \times N$). While the relationship matrix is not a correlation matrix per se, note that we can compute the sample correlation matrix between the genotypes as

$$\mathbf{C} = \frac{1}{(N-1)}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \tag{2.17}$$

where $\mathbf{X}$ is a $(N \times M)$ matrix with the (i, k) elements equal to the normalized values

$$x_{ik} = \frac{n_{ik} - \overline{n}_{.k}}{\sqrt{S^2_{.k}}}.$$

Here $\overline{n}_{.k}$ is the usual sample mean for SNP k and ${S_{.k}}^2$ is equal to the usual sample variance estimator $(1/N-1)\sum_{i=1}^{N}(n_{ik} - \overline{n}_{.k})^2$. Note the similarity between this x_{ik} and z_{ik} in (2.7) which is used in the calculation of the estimated relationship matrix $\hat{\mathbf{K}}$. In fact the estimate of the mean $\overline{n}_{.k}$ is exactly equal to $2\hat{p}_{.k}$, and $\sqrt{2\hat{p}_{.k}}(1-\hat{p}_{.k})$ will be close (if HWE is not greatly violated) to $\sqrt{S^2_{.k}}$. This implies that the estimated relationship matrix (computed for all pairs of a total of N subjects) is "almost" equal to the $N \times N$ matrix $(1/M)\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Moreover it is easy to show that if α_l is the lth eigenvector of $(1/(N-1))\mathbf{X}^{\mathrm{T}}\mathbf{X}$, the lth eigenvector of $(1/M)\mathbf{X}\mathbf{X}^{\mathrm{T}}$ will be equal to a constant times $\mathbf{X}\alpha_l$. Because of this, the eigenvectors of the relationship matrix while computed using the $z_{i,k}$ are generally almost proportional to the principal components computed using the eigenvectors of the correlation matrix computed from the $x_{i,k}$.

Note that if (as typical in a GWAS study) we have a larger number, M, of SNPs (several hundred thousand) genotyped than we have individuals, M (several thousand), then it may be much more computationally efficient to extract eigenvectors from the $N \times N$ relationship matrix than from the $M \times M$ correlation matrix. Given these considerations, it is reasonable to accept the slight abuse of terminology of Price et al. and refer to the eigenvectors of $\hat{\mathbf{K}}$ as the principal components of the genotype data.

2.9.1 Display of Principal Components for the HapMap Phase 3 Samples

The HapMap project expanded the number of populations with genotypes available by (in phase 3) genotyping a total of 1,184 subjects from 11 different distinct population samples using a combination of the Affymetrix 6.0 and Illumina 1M SNP arrays, each of which genotyped about one million SNPs with approximately 400,000 SNPs genotyped using both arrays. We illustrate the power of principal components analysis to extract the main features of historical separation that arose between populations, as groups diverged from their ancestral origins, and more recently admixture between long-separated groups, by use of a sample of SNPs

from these data. We illustrate the process here of using the HapMap website in order to read into *R* a sample of ~20,000 randomly chosen SNPs from each of the 11 HapMap phase 3 populations, perform the calculation of the relationship matrix and its principal components (e.g., eigenvectors), and then display simple plots of these data. Because we will only be using a small sample of the HapMap genotypes, we can do all of the statistical calculations and plotting reasonably quickly in *R*; for larger datasets, there are several other choices depending on the expertise of the statistician, for example, the stand-alone program EIGENSTRAT, introduced by Price et al. computes principal components, but then these must be read into other programs (such as *R*) for plotting, filtering of results, summarization, etc. The statistical package SAS and its principal components procedure PRINCOMP can also be used to compute and display eigenvectors of $\hat{\mathbf{K}}$. These programs are effective in computing eigenvectors for quite large matrices (~10,000 subjects or more) on a modern desktop computer with a large amount of main memory (8–16 Gb).

For the illustrative example here, the simplest way to retrieve data from the HapMap website is as a PLINK format file and to use PLINK (see Computer Appendix) to extract a random sample of SNPs. The following steps were followed:

Download HapMap phase 3 data set, *hapmap3_r1_b36_fwd.qc.poly.tar.bz2*, from the HapMap website ftp site *ftp://ftp.ncbi.nlm.nih.gov*. This file, which contains separate files for each of the 11 contributing population samples, can be decompressed using the linux command *bunzip* or the Windows program *winzip*. The 11 HapMap 3 populations each have files of type *.ped which contain genotypes and *.map files which contain SNP information.

Using PLINK, merge these 22 files together, make a random selection of approximately 20,000 SNPs to use in the PCA analysis, and perform the PCA analysis and plotting using *R*. To do so, we first run the following PLINK commands:

```
plink --file hapmap3_r2_b36_fwd.ASW.qc.poly --merge-list allfiles.txt --make-bed --out
hapmap3_allpops
plink --bfile hapmap3_allpops --recodeA --thin 0.02 –maf 0.05 --filter-founders --out
hapmap3_snp_sample
```

The first command refers to a list of all the files (*allfiles.txt*, created by the user) to be merged with the first file (ASW). The *allfiles.txt* file lists the remaining 10 genotype files to be merged starting with the first one as listed on the command line. It contains the list:

```
hapmap3_r2_b36_fwd.CEU.qc.poly.ped hapmap3_r2_b36_fwd.CEU.qc.poly.map
hapmap3_r2_b36_fwd.CHB.qc.poly.ped hapmap3_r2_b36_fwd.CHB.qc.poly.map
hapmap3_r2_b36_fwd.CHD.qc.poly.ped hapmap3_r2_b36_fwd.CHD.qc.poly.map
hapmap3_r2_b36_fwd.GIH.qc.poly.ped hapmap3_r2_b36_fwd.GIH.qc.poly.map
hapmap3_r2_b36_fwd.JPT.qc.poly.ped hapmap3_r2_b36_fwd.JPT.qc.poly.map
hapmap3_r2_b36_fwd.LWK.qc.poly.ped hapmap3_r2_b36_fwd.LWK.qc.poly.map
hapmap3_r2_b36_fwd.MEX.qc.poly.ped hapmap3_r2_b36_fwd.MEX.qc.poly.map
hapmap3_r2_b36_fwd.MKK.qc.poly.ped hapmap3_r2_b36_fwd.MKK.qc.poly.map
hapmap3_r2_b36_fwd.TSI.qc.poly.ped hapmap3_r2_b36_fwd.TSI.qc.poly.map
hapmap3_r2_b36_fwd.YRI.qc.poly.ped hapmap3_r2_b36_fwd.YRI.qc.poly.map
```

These files being merged are named according to the population descriptors for the 11 HapMap phase 3 groups: See Table 2.5.

The output of the second PLINK command is the file, hapmap3_snp_sample.raw, which is formatted in a way that the *R* command, *read.table*, can read directly. The *–recodeA* subcommand codes the SNP alleles as equal to 0, 1, or 2 copies of the minor allele with NA being the missing value indicator. The *–maf 0.05* command removes SNPs which have minor allele frequency less than 5 %, and the *–filter-founders* command indicates that only genotype data for the total of 988 founders (i.e., dropping offspring from any parent-offspring trios) is to be included.

With the addition of one other file (*IDgroup.txt* which contains the ID record for and group affiliation (ASW, CEU, etc. for each subject), we can now perform a simple PCA analysis in *R*.

Read data into *R* and compute the estimated relationship matrix for all *X* subjects:

```
>Calc_k<-function(x){
>       nsubj<-dim(x)[1]
>       nsnps<-dim(x)[2]
>       cfreq<-colMeans(x,na.rm=TRUE)/2
>       y<-t(t(x)-cfreq*2)
>       y<-ifelse(!is.na(y),y,0)
>       cv<-2*cfreq*(1-cfreq)
>       K<-t(t(y)/cv)%*%t(y)/nsnps
>       K
>}
>
>pca_snps<-read.table("hapmap3_snp_sample.raw",header=T)
>
>IDgrp<-read.table(file="IDgrp.txt")
>grp<-as.character(IDgrp[,2])
>ID<-as.character(IDgrp[,1])
>
>ncols<-dim(pca_snps)[2]
>K<-Calc_k(pca_snps[1:nsubj,7:ncols]) #the first 6 columns do not contain genotypes
```

Next we compute eigenvectors and define and run a function *plotPC* used to display the results:

```
> e<-eigen(K)
>
> # make up a function to do this  ##### terrible code USE matplot instead
> #
> #
> plotPC<->function(e,GRP=c("ASW","CHB","CHD","GIH","JPT","LWK","MEX","MKK","TSI","YRI"),
>       COLOR=c("red","blue","green","yellow","orange","black","purple","pink","gray","violet"
>       ,"cyan>"),d1=1,d2=2,xl=c(-0.05,0.05),yl=c(-0.05,0.05),gp=grp)
>       {
>       matplot(rbind(e$vectors[gp=GRP[1],1]),e$vectors,xlim=xl,ylim=yl,col=COLOR,main="
>       ",xlab=paste("EV",d1,sep=""),ylab=paste("EV",d2,sep=""))
>          lr<-length(rest)
>          if (length(rest) >= 1) {
>                  for (i in 2:length(rest))
>                   # add additional populatons
>                  lines(e$vectors[gp==rest[i-1],d1],e$vectors[gp==rest[i-1],d2],
>                  xlim=xl,ylim=yl,col=restc[i-1],type="p")
>       }
> }
> # run the plotting function
> plotPC(e)
```

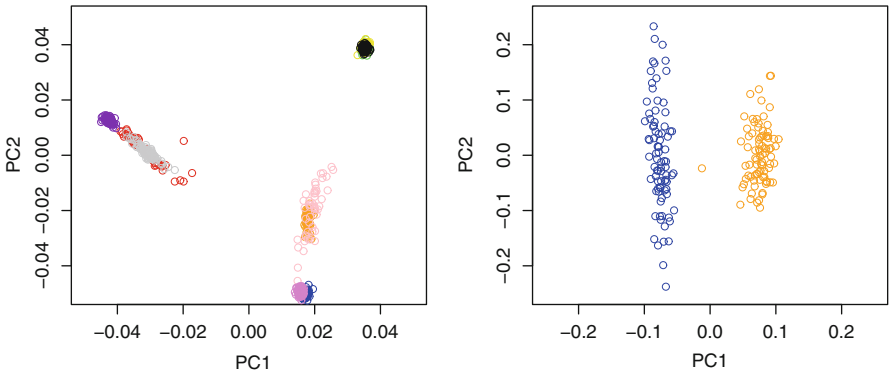

Fig. 2.5 Principal components plot for the HapMap phase 3 data. R code for this plot is given in the text. The *colors* correspond to group membership; see the *R* function plotPC, defined in the text, for color coding. The *left side plot* shows all 11 groups plotted, while the *right side plot* is only the Japanese and Chinese samples

The results of this first call to the function *plotPC* are given in Fig. 2.5a.

In Fig. 2.5a the main visible features are the clustering of the three continental groups (European, Asian, and African) as well as evidence of admixture for several groups including the African Americans (ASW) and the Mexican Americans (MEX). We can look in more detail at the two East Asian groups, Japanese from Tokyo (JPT) and Chinese from Beijing (CHB), by recomputing the K matrix that pertains to these two groups, recalculating the eigenvectors, and again calling the *plotPC* function:

```
>grpJC<-(grp=="JPT")|(grp=="CHB")
>KJC<-Calc_k(pca_snps[grpJC,7:ncols])
>ejc<-eigen(KJC)
>gJC<-grp[grpJC]
>
>plotPC(ejc,gp=gJC,GRP="JPT",d1=1,d2=2,COLOR="orange",rest=c("CHB"),restc=("blue"),xl=c(-
>.25,.25),yl=c(-.25,.25))
>
```

The results are shown in Fig. 2.5b. Now the first eigenvector is sensitive to differences in SNP frequencies between the Japanese samples, compared to the Chinese samples. This plot shows that the principal components method can make more subtle distinctions than just the detection of large-scale continental variation. This Japanese group of samples appears to derive from China given the overall similarity of the two groups seen in Fig. 2.5a, but remains partly distinguishable as in Fig. 2.5b. The observed differences are probably related to founder effects and random genetic drift due to the relative isolation of Japan from the mainland of Asia over a number of generations [48].

Principal components analysis is of course interesting since it says much about historical relationships among population groups; it has also been stressed here

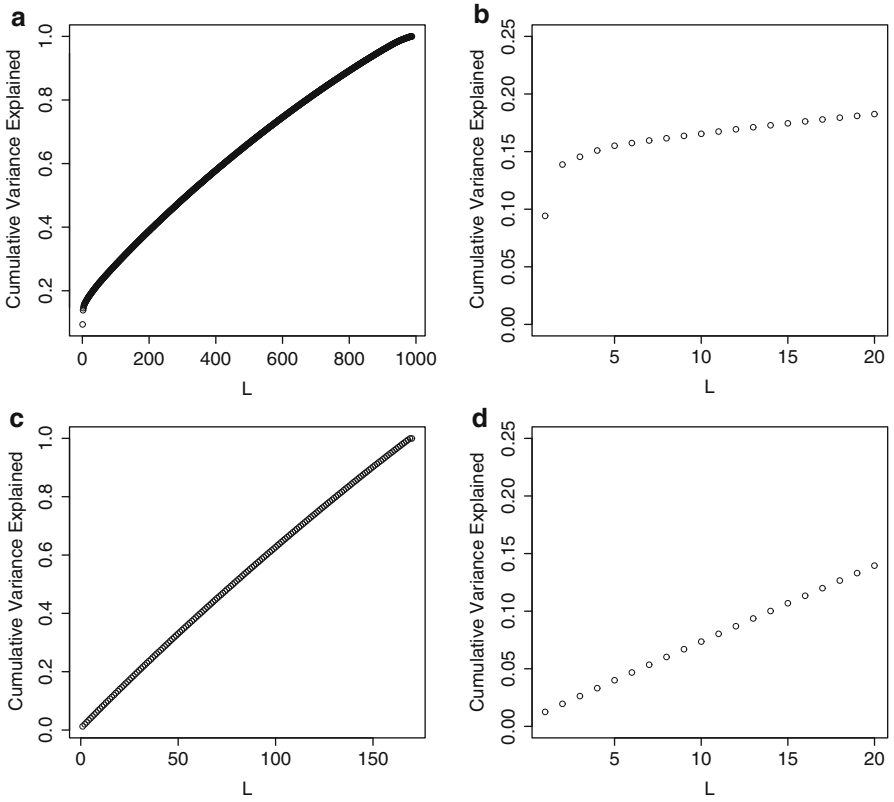

Fig. 2.6 Plots of the total variation explained by eigenvectors 1, . . ., L for (**a**, **b**) the entire HapMap phase 3 dataset and (**c**, **d**) for the Japanese (JPT) and Chinese (CHB)

because of its utility as a tool in association studies as will be described in later chapters. Principal components can also be used as a QC tool in checking that batches of genotype data are not being disturbed by subtle factors (such as plate specific errors and DNA quality issues). Principal components are sensitive to each of population structure, relatedness between subjects, and to certain patterns of linkage disequilibrium, as well as such QC problems.

It is worth saying also that while principal components can distinguish groups that are quite closely related historically, the total amount of SNP variation that is explained by the first few leading principal components—i.e., the portion of SNP variation that can be directly related to racial/ethnic origins, is actually quite small. Figure 2.6 plots the cumulative percentage (2.16) of SNP variation explained by the first L eigenvectors for both the total HapMap phase 3 population and for the Japanese and Chinese separately. About 15 % of the total variation in the entire HapMap sample and much less (~2 %) of the variation of the SNP data for the Japanese and Chinese is explained by the two eigenvectors plotted in Fig 2.6a, b.

This illustrates the general concept that racial groups are ultimately much more similar than they are different, or more precisely that within-group genetic variation dwarfs between-group variation.

2.10 Chapter Summary

The chapter has presented a brief survey of topics related to population and quantitative genetics that are relevant to the motivation for and the design and analysis of large-scale genetic association studies. The concepts of linkage disequilibrium, population heterogeneity and relatedness, and the induced covariance between markers caused by population heterogeneity, as well as the techniques of principal component analysis, described here, are all followed up in the later chapters in this book.

Homework/Projects

1. If the vector of allele counts $\mathbf{n_1} = (n_{11}, n_{21}, \ldots, n_{N1})'$ for a given variant has covariance matrix $2p_1(1 - p_1)\mathbf{K}$ and that another variant with allele counts $\mathbf{n_2} = (n_{12}, n_{22}, \ldots, n_{N2})'$ independent of $\mathbf{n_1}$ which has covariance matrix $2p_2(1 - p_2)\mathbf{K}$:
   (a) Show that the sum of these two variants, $\mathbf{s} = \mathbf{n_1} + \mathbf{n_2}$, will have covariance matrix $(2p_1(1 - p_1) + 2p_2(1 - p_2))\mathbf{K}$.
   (b) Show that the correlation matrix of $\mathbf{s}$ is the same as the correlation matrix of $\mathbf{n_1}$ (or for that matter $\mathbf{n_2}$).
   (c) Define a *polygene*, $g_i = \sum_{j=1,\ldots,M} w_j n_{ij}$, as a weighted sum of M independent alleles each with the same covariance structure (covariance between subjects i and j) as above. Show that the covariance structure of the polygene is equal to $\gamma^2\mathbf{K}$ with $\gamma^2 = \sum_{j=1,\ldots,M} w_j{}^2 2p_j(1 - p_j)$
2. Give details of the calculations for several other probabilities in Table 2.2 following the method used for the (Aa, Aa) genotype.
3. For first cousins (where the grandparents are unrelated to each other):
   (a) Compute the probabilities z_0, z_1, and z_2 from first principles.
   (b) Suppose the first cousins marry and have two children, what is each offspring's inbreeding coefficient and the probabilities z_0, z_1, and z_2 between them?
4. Show using Table 2.2 that
$$\mathrm{Cov}(n_1, n_2) = (z_1 + 2z_2)p(1 - p).$$
5. Section 2.1.3: Suppose that DNA for a random sample of pairs of siblings is obtained and all DNA samples are genotyped for a polymorphism. How does

(2.1) need to be modified to give a confidence interval for the allele frequency of an allele A?

Hint: first find the variance of the allele frequency estimate for each sibling pair and then the average of this estimate over all pairs.

6. Section 2.1.4, inbreeding:

   (a) Show that if an individual's parents are related with coefficient of kinship equal to h then if n_A is the count of allele A for a given polymorphism $\{aA\}$ with frequency of A equal to p, the variance of n_A is $2p(1-p)(1+h)$.
   (b) Show also that the predicted fraction of heterozygotes equals $2p(1-p)(1-h)$ in the same situation. This gives a definition of F_{st} for an inbred population, namely, as the average kinship between individuals in the population, which is estimated by the reduction of heterozygotes away from HWE.

   See Balding and Nichols' papers [49–51] for more details.

7. Section 2.5.1, relatedness: compute the variance of $\frac{1}{M}\sum_{\ell=1}^{M} z_{i\ell}^2$ under assumptions of (1) Hardy–Weinberg equilibrium for all SNPs (implying that this quantity has expected value equal to one), (2) that all variants are independent (not in LD) of each other, and (3) that all have the same allele frequency p. Compute the variance under the same conditions of the suggested estimator (Yang et al. [9])

$$1+\frac{1}{M}\sum_{\ell=1}^{M}\frac{n_{i\ell}^2-(1+2p_\ell)n_{i\ell}+2p_\ell^2}{2p_\ell(1-p_\ell)}.$$

   Hint: compute from first principals the mean and second moments of ${z_{i\ell}}^2$ under HWE by summation over its distribution, i.e., ${z_{i\ell}}^2$ takes value $(0-2p_\ell)^2/(2p_\ell(1-p_\ell))$ with probability $(1-p_\ell)^2$, value $(1-2p_\ell)^2/(2p_\ell(1-p_\ell))$ with probability $2p_\ell(1-p_\ell)$, and $\frac{(2-2p_\ell)^2}{2p_\ell(1-p_\ell)}$ with probability ${p_\ell}^2$.

8. Section 2.9, principal components. Why are the eigenvalues of a covariance matrix $\mathbf{\Sigma}$ considered to be equivalent to variances explained by eigenvectors, and why is (2.16) interpreted as the fraction of variance explained by the first L leading eigenvectors?
9. Section 2.7.4: Show that for a mixed population, Z, treated as a single population containing two subpopulations, X and Y, with proportion m and $(1-m)$, respectively, that the disequilibrium between any two markers A and B in the mixed population is equal to

$$\delta_Z = m\delta_X + (1-m)\delta_Y + m(1-m)\left(p_{A,X}-p_{A,Y}\right)\left(p_{B,X}-p_{B,Y}\right),$$

   with δ_X, δ_Y, and δ_Z as the disequilibrium parameters in populations, X, Y, and Z, respectively, and allele frequencies of markers A and B as $p_{A,X}$, $p_{B,X}$, $p_{A,Y}$, and

$p_{B,Y}$ in populations X and Y, respectively. Hint: see reference [52]. Note that this implies that markers which are unlinked in the originating populations ($\delta_X = 0$, and $\delta_Y = 0$) will be linked in a stratified populations only if the allele frequencies of both alleles are different.

10. Section 2.4.1.1: Based on the Balding–Nichols model used in that section give a formula for H_t and H_s and hence F_{st} when F remains the same for each population t, but the sample size in each population, N_t is different.
11. Section 2.6.3, selection and allele frequency distribution. Consider a cancer like breast cancer and assume that the only reduction in reproductive fitness related to occurrence of breast cancer is due to early mortality (e.g., mortality in the childbearing years from the cancer). We are interested in the following question: if a genetic variant of 10 % frequency increases relative risk of breast cancer mortality by 20 % per allele proportionately over a woman's lifetime, what value of σ in (2.14) would this correspond to, i.e., what reduction of fitness does this imply? What data would be needed in order to address this question?
12. Section 2.6.1, having two sexes raises an obvious question. Is there a most recent common female ancestor and a most recent common male ancestor? If so did they know each other? For a discussion, see http://scienceblogs.com/authority/2007/07/17/adam-eve-and-why-they-never-go/.

2.11 Data and Software Exercises

Here we assume that the JAPC and LAPC data mentioned in Chap. 1 are available to the student, either through a dbGaP request or otherwise. If other GWAS data are available to the student, these can in most cases be used instead.

1. Using available GWAS data (e.g., JAPC and LAPC) as well as the HapMap data described above, perform principal components analysis on the combined data. This consists of several steps:
   (a) Develop a list of SNPs that are genotyped in both the GWAS data and the HapMap data described above (hint: use the bim files to manipulate them in SAS or *R*) to form a list of the intersection of SNPs between all these files.
   (b) Sample perhaps 10,000 or so markers from the intersection above (SAS or *R*).
   (c) Pull out SNP data from each file (using PLINK commands). For HapMap do not use offspring who have parents with genotypes (PLINK command).
   (d) Merge the data for the 10,000 common markers into a single PLINK file (PLINK commands).
   (e) Recode this file to 0, 1, 2 coding (count of the number of minor alleles carried for each SNP) using PLINK commands.
   (f) Read this file into *R* and modify/run the *R* programs given above to extract principal components of the data and display these data as done above for the HapMap samples.

(g) Make a special run just on the JAPC data in combination with the JPT HapMap population. Describe the finer structure in the JAPC data; see [48] for a survey of population structure in Japan.

2. EIGENSTRAT for principal components estimation. We have shown a simple *R* program that can estimate the relationship matrix **K** using the formulae (2.8) and (2.9). This program will work reasonably well for a few thousand markers and few hundred subjects. A much more efficient program that can handle hundreds of thousands of markers and tens of thousands of subjects is EIGENSTRAT [53] available from http://genepath.med.harvard.edu/~reich/Software.htm. Review the use of this program.
3. STRUCTURE, Pritchard et al.[54]. This program is specifically designed to identify hidden structure in a population based upon SNP data and estimate population membership or admixture fraction for each individual in a study. The results of the STRUCTURE analysis (an individual's percent admixture from different populations) is often more interpretable than using principal components and has been used in many analyses. The program can be downloaded from URL http://pritch.bsd.uchicago.edu/structure.html.

(a) Try running this program on samples of the HapMap data described above. Estimate admixture fractions for individuals in the ASW group (African ancestry in Los Angeles CA).
(b) Using the JAPC data, does this program estimate the fine structure visible by contained in this population?

References

1. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*, 559–575.
2. Choi, Y., Wijsman, E. M., & Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology, 33*, 668–678.
3. Wakeley, J. (2009). *Coalescent theory, an introduction*. Greenwood Village, CO: Roberts and Company.
4. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*, 904–909.
5. Balding, D., & Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic locus and its implications for investigating identify and paternity. *Genetica, 3*, 3–12.
6. Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics, 15*, 323–354.
7. Chen, G. K., Millikan, R. C., John, E. M., Ambrosone, C. B., Bernstein, L., Zheng, W., et al. (2010). The potential for enhancing the power of genetic association studies in African Americans through the reuse of existing genotype data. *PLoS Genetics, 6*, e101096.
8. Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., et al. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *American Journal of Human Genetics, 73*, 612–626.

9. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics, 42*, 565–569.
10. Astle, W., & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science, 24*, 451–471.
11. Nordborg, M. (2008). Coalescent theory. In D. J. Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of statistical genetics* (3rd ed., pp. 843–877). New York: Wiley.
12. Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics, 18*, 337–338.
13. Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., et al. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature, 445*, 82–85.
14. Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*, 1516–1517.
15. Lander, E. S. (1996). The new genomics: Global views of biology. *Science, 274*, 536–539.
16. Chakravarti, A. (1999). Population genetics-making sense out of sequence. *Nature Genetics, 21*, 56–60.
17. Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics, 17*, 502–510.
18. Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: Common disease-common variant. . .or not? *Human Molecular Genetics, 11*, 2417–2423.
19. Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America, 24*, 253–259.
20. Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology, 3*, 87–112.
21. Slatkin, M., & Rannala, B. (1997). The sampling distribution of disease-associated alleles. *Genetics, 147*, 1855–1861.
22. Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*, 56–65.
23. Wright, S. (Ed.). (1949). *Adaptation and selection*. Princeton, NJ: Princeton University Press.
24. Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics, 69*, 124–137.
25. Griffiths, R. C., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology, 3*, 479–502.
26. Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science, 310*, 321–324.
27. Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics, 4*, 587–597.
28. Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage disequilibrium in the human genome. *Nature, 411*, 199–204.
29. McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science, 304*, 581–584.
30. Lewontin, R. (1964). The interaction of selection and linkage. I.general considerations: Heterotic models. *Genetics, 49*, 49–67.
31. Thomas, D. C. (2004). *Statistical methods in genetic epidemiology*. Oxford: Oxford University Press.
32. Gulcher, J., & Stefansson, K. (1998). Population genomics: Laying the groundwork for genetic disease modeling and targeting. *Clinical Chemistry and Laboratory Medicine, 36*, 523–527.
33. Editorial Board. (1998). Genome vikings. *Nature Genetics, 20*, 99–101.
34. Bhattacharjee, S., Wang, Z., Ciampa, J., Kraft, P., Chanock, S., Yu, K., et al. (2010). Using principal components of genetic variation for robust and powerful detection of gene-gene

interactions in case-control and case-only studies. *American Journal of Human Genetics, 86*, 331–342.
35. Hao, K., Schadt, E. E., & Storey, J. D. (2008). Calibrating the performance of SNP arrays for whole-genome association studies. *PLoS Genetics, 4*, e1000109.
36. Barrett, J. C., & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics, 38*, 659–662.
37. Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., et al. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science, 314*, 1930–1933.
38. Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., & Donnelly, P. (2005). A haplotype map of the human genome. *Nature, 437*, 1299–1320.
39. Kitts, A., & Sherry, S. (2002). The single nucleotide polymorphism database (Dbsnp) of nucleotide sequence variation. In J. McEntyre & J. Ostell (Eds.), *The NCBI handbook [Internet]*. Bethesda, MD: National Center for Biotechnology Information.
40. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*, 1061–1073.
41. Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology, 9*, 640.
42. Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nature Genetics, 33*(Suppl), 266–275.
43. Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.
44. Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics, 17*, 763–774.
45. Rao, P. K., & Li, Q. (2009). Principal component analysis of proteome dynamics in iron-starved mycobacterium tuberculosis. *Journal of Proteomics and Bioinformatics, 2*, 19–31.
46. Chang, E. T., Lee, V. S., Canchola, A. J., Dalvi, T. B., Clarke, C. A., Reynolds, P., et al. (2008). Dietary patterns and risk of ovarian cancer in the California teachers study cohort. *Nutrition and Cancer, 60*, 285–291.
47. Preisendorfer, R. W. (1988). *Principal components analysis in meteorology and oceanography*. Amsterdam: Elsevier.
48. Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M., et al. (2008). Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: Effects on population-based association studies. *American Journal of Human Genetics, 83*, 445–456.
49. Balding, D. J., & Nichols, R. A. (1994). DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International, 64*, 125–140.
50. Nichols, R. A., & Balding, D. J. (1991). Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity Edinburgh, 66*(Pt 2), 297–302.
51. Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica, 96*, 3–12.
52. Chakraborty, R., & Smouse, P. E. (1988). Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. *Proceedings of the National Academy of Sciences of the United States of America, 85*, 3071–3074.
53. Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics, 11*, 459–463.
54. Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*, 945–959.

http://www.springer.com/978-1-4614-9442-3

Design, Analysis, and Interpretation of Genome-Wide Association Scans

Stram, D.O.

2014, XV, 334 p. 39 illus., Hardcover
ISBN: 978-1-4614-9442-3