8

Marker-based Estimation of Relatedness

Version 13 August 2022

The machinery of quantitative genetics relies heavily on having sets of known relatives. This is not a problem in controlled experiments or when one has access to pedigrees, but these are special settings, and their requirement significantly constrains the universality of these methods. Our recent ability to score tens of thousands of DNA markers rapidly, cheaply, and accurately has opened up essentially any population to a quantitative-genetic analysis, as we can estimate relatedness from this marker information alone.

The idea behind using marker information to estimate the relationship between a pair of individuals is straightforward in concept: relatives should share alleles that are identical by descent (IBD; Chapter 7). Alas, in most settings we cannot *directly score* IBD alleles, but rather must *infer* the fraction of IBD alleles from information on alike in state (AIS; Chapter 7) alleles. IBD alleles are AIS, but the converse is not necessarily true. The amount of IBD sharing can be estimating using either an excess of AIS alleles relative to some neutral expectation or as a correlation between AIS alleles shared by individuals. Implicit in this idea is the notion of some **base population**, from which deviations are measured. Given a randomly mating and unstructured base population, estimated allele frequencies can be used to obtain the expected AIS similarity between two unrelated individuals. The challenging issue with marker data is not simply declaring that two individuals are related, but rather in estimating their *actual degree of relationship*, typically by estimating the coefficient of coancestry, Θ (Chapter 7).

One of the earliest uses of markers was in paternity testing. A famous case from 1943 was the paternity suit of the actress Joan Barry, who claimed that her child was fathered by fellow actor Charlie Chaplin. Chaplin's blood type was *AB*, while both Joan and her child were both *O*. In a tribute to the scientific acumen of juries, Chaplin was declared to be the father. The early crude level of resolution for paternity and forensic issues offered by blood-type testing (**serology**) has since been replaced by highly accurate assessment using DNA markers. In particular, paternity testing and crime scene analysis (whether a biological sample came from a specified person) is now based (in the US) on just 20 highly polymorphic loci (the CODIS markers), which (roughly speaking) tend to have around ten alleles per locus, all at roughly equal frequencies. Hence the chance of a random individual matching a crime scene DNA profile is (very crudely) on the order of $[2(1/10)^2]^{20} \simeq 10^{-34}$. With a small number of such highly polymorphic markers, we can also assess whether two individual are very close relatives (parent-offspring, full versus half sibs). Detecting more distant relationships requires significantly more markers, as the expected fraction of IBD alleles scales as $(1/2)^t$, where t is the total number of generations that separate the individuals from their shared ancestor (Chapter 7). A further complication is that there is considerable variation about this expected value, such that many distant relatives will share *no* IBD regions.

Even with a known pedigree, markers are still very useful. First, the veracity of the pedigree can be checked, as **incorrectly ascertained paternities** are not uncommon, even in systems with apparently strong control over matings. For example, Visscher at al. (2002) estimated a sire error rate of $\sim$10% for UK dairy cattle, despite the very widespread use of artificial insemination, while Leroy et al. (2012) found rates of 1–9% for dogs, 1–10% for sheep, and 4% for a French cattle population. Recording errors, as well as the ingenuity of organisms searching for mates, should never be underestimated! Likewise, especially in natural populations, it may be unclear if the offspring from a female are full- or half-sibs. Such close-relative uncertainties can be easily dealt with using a modest number of highly

polymorphic markers. Second, a more subtle issue (first discussed in Chapter 7) is the notion of *expected versus realized relatedness*. Θ values computed from pedigrees represent *expected values*, while the *realized* or *actual* relatedness is a random variable distributed around this expected value. For example, while the expected value of Θ for noninbred full sibs is $1/4$ (all pairs of full sibs in a pedigree are assigned this value), the actual relationship for any pair of full sibs can deviate, potentially quite significantly, from 0.25 (Figure 7.5). As we will see in later chapters, this variation around the expected value can be powerfully exploited in a variety of settings. Accurate estimates of such realized relatedness requires *at least* a few thousand markers.

Our presentation is as follows. We start with some general comments about marker classes and their associated assays. Next, we examine the use of markers to assign individuals into a few discrete **relationship categories**, such as paternity assessment and whether offspring in a family are half- or full-sibs. These approaches require only a modest number of markers, and formed the basis for much of the early work on marker-based estimates of relatedness. As marker density greatly increased, so did the ability to detect much more distant relatives. Rather than trying to assign pairs of such relatives into discrete relationship categories, we instead describe their associations using continuous **relatedness** measures, such as Θ_{xy}, f_x, and Δ_{xy}. We first consider a few foundational issues of relatedness measures, such as setting a base population, IBD probability versus allelic correlation frameworks, and the variance in relatedness before considering estimation strategies.

MOLECULAR MARKERS AND GENOMICS

A **marker** is simply a locus with scorable genetic variation. Markers are used in quantitative genetics for *estimating relatedness* (the focus of this chapter), an idea that can be extended to marker-based BLUP and genomic prediction (Chapters 31 and 32). Their other important use is in providing *linkage information* for either linkage-based QTL mapping (such as with a line cross) or association studies (GWAS) that exploit population-level linkage disequilibrium (Chapters 17–21). Our discussion here of different marker systems will consider their impact on both of these objectives.

The ideal class of markers would be: (i) codominant (heterozygotes and homozygotes can be distinguished), (ii) easy and cheap to score, (iii) generally not impacting any trait, (iv) numerous (the more markers the better), (v) scattered throughout the genome, and (vi) easily obtainable (a set of markers can be generated for *any* population of interest). As we detail shortly, molecular markers (such as SNPs and STRs) nicely satisfy most of these requirements.

In the classical genetics era, markers were usually **morphological**, impacting the phenotype, such the recessive *hairy eye* allele and its dominant normal (wild type) allele in *Drosophila melanogaster*. As in this example, many of these morphological markers are dominant, complicating their use. Such markers were used in very early (and crude) QTL mapping studies in *Drosophila* (Chapter 18). There were also a few early molecular markers, such as blood groups (where alleles *A* and *B* are dominant to *i*, with *ii* being type *O*). The first widespread molecular markers used were **allozymes** (also known as **isozymes**), protein variants detected by differences in migration on starch gels in an electric field. The resulting band positions are then revealed using protein-specific stains (Hubby and Lewontin 1966; Lewontin and Hubby 1966). Since the late 1960s, this class of markers has been extensively applied to a variety of population-genetic problems (Lewontin 1974; Charlesworth et al. 2016), and were used in some of the early QTL mapping experiments (Chapter 18). As methods for evaluating variation directly at the DNA level became more widely available during the mid-1980s, DNA-based markers largely replaced allozymes. Given that a set of markers tends to be population-specific (or at least largely species-specific), a key feature beyond the *scoring* of markers is in *generating them in the first place*. Namely, given a novel population or species without any existing marker information, a new set of can markers be quickly and cheaply generated. Modern genomics solves this problem.

Properties of DNA Variation

Before briefly reviewing a few legacy systems for scoring markers based on DNA variation (RFLPs, RAPDs, etc.), it is worthwhile to consider the different types of variation that one might see a population sample of fully-sequenced genomes. The key properties of any such DNA-level variation can be divided into two categories. First, properties that are intrinsic to a class of markers (their stability, genome abundance, levels of polymorphism, and potential restrictions in their genomic coverage, such as coding versus noncoding regions). Second, properties that arise as a result of the assays used to score such makers; most notably, whether heterozygotes can be distinguished from homozygotes (whether the markers are codominant when scored by a particular assay). We group sources of DNA variation as follows:

1. Single nucleotide polymorphisms (**SNPs**). These reflect base-pair differences (polymorphisms) at a specific single nucleotide in our sample. While in theory up to four alleles could be segregating at any single site (representing the four potential DNA bases), most SNPs are biallelic. For these, we code the two alleles as 0 or 1, with the allele coded as 1 denoted as the **reference allele**, which is usually taken as the more frequent allele. The **minor allele frequency**, or **MAF**, denotes the frequency of allele 0. SNP alleles are stable, with mutation rates for nuclear-encoded genes on the order of no greater than 10^{-9} per nucleotide per generation, and often much less (WL Table 4.2). SNPs are abundant and widely dispersed throughout the genome, although they tend to be underrepresented in genomic regions of greatly suppressed recombination (such as around centromeres), likely reflecting the action of both positive and negative selection at linked sites (WL Chapter 8). Their one (minor) deficiency is that a single biallelic SNP is usually not very informative, with a maximum heterozygosity of $1/2$ (the maximum value occurs at $p = 1/2$). Further, the minor allele for neutral SNPs is often expected to be rare (WL Equation 2.34), so that most SNPs have a heterozygote frequency far less than 50%. This limitation is easily overcome as one can score a vast number of SNPs, obtaining, in aggregate, a significant amount of polymorphic information. Additionally, there is an ascertainment bias in how SNP are generally chosen, such that **common SNPs** tend to be greatly overrepresented. Common SNPs are defined as having intermediate allele frequencies: MAFs of greater than one to five percent, giving heterozygosities of at least two to ten percent. Finally, most SNPs are regarded as **anonymous DNA markers** (usually—but not always—expected to have no direct impact on a phenotype).

2. Microsatellites. A second common category of DNA variation is the tandem repetition of small blocks (**motifs**) of sequence, such as $AGGAGGAGGAGG$. Such **tandem arrays** are often referred to **satellites** (a holdover from the early days of DNA analysis, when repetitive DNA sequences often formed a distinct band, or *satellite*, when extracted DNA was centrifuged in a salt gradient). The genotype of a tandem array is typically scored as its number of repeats, so that a 4,7 heterozyote contains one 4-unit repeat and a 7-unit repeat in the same location on the two homologous chromosomes. The size of the repeat unit forming the array can range from a few bases to thousands of bases, leading to microsatellites, minisatellites, and so on. The most commonly scored are those involving repeating units of just a few bases, and these are called **simple sequence repeats** (**SSRs**) or **simple tandem repeats** (**STRs**). Their major advantage is that any given SSR usually consists of a number of alleles, often of roughly equal frequency. For this reason, the markers used in forensic DNA are well-chosen SSRs that typically contain around 10 alleles of roughly equal frequency, for a heterozygosity of around 90% per SSR (with k equally frequent alleles, the heterozygosity is $[1 - 1/k]$). As a result, just a few SSRs can contain as much polymorphic information as a dozens, or even hundreds, of SNPs. The highly polymorphic nature of SSRs arises because of their very high mutation rates (on the order of 10^{-2} to 10^{-4} per generation). This makes ***SSR alleles much more***

unstable than SNP alleles, resulting in signals from distant relationships being partly obscured by mutations changing the allelic state. As a result, SSRs are excellent for close relatives, but become increasingly problematic as the most recent common ancestor for a pair of relatives moves further back in time. The same concern arises when using SSRs for linkage information when the common ancestor is several generations back. An association between a particular SSR allele and a nearby QTL can be broken by *mutation* rather than *recombination*. For example, if initally a *5* allele at an SSR and a *Q* allele at a QTL are together in a single linkage block (**haplotype**), an SSR mutation can change this to a *6 – Q* haplotype. Hence, SSRs may be suitable for linkage mapping (one or a few generations following a cross), but not for association analysis (which uses signals from common relatives that are potentially hundreds of generations back; Chapter 20).

3. Insertion and deletions (**indels**). While the most widely used DNA markers in quantitative genetics are either SNPs or SSRs, the genome is a highly dynamic structure, and our final three classes concern such **structural variation**. One common feature are insertions and deletions of various sizes. As with satellites, the scale of such insertions and deletions can vary dramatically, although the larger the indel, the more likely it is to be deleterious (and thus rare in a sample). Indels of just a few bases are rather common, as are indels generated by mobile genetic elements, which are abundant in most genomes. Given that such elements can jump into a gene (disrupting its function) and often contain long-range acting *cis* regulatory factors, mobile elements have the potential to influence the expression of nearby QTLs. Mobile elements can also self-excise (jump-out), either cleanly, or leaving some residual sequence behind.

4. Copy number variation (**CNV**). Potential contributors to quantitative-genetic variation are differences in the number of copies of a particular gene among individuals within a sample. Such copy number variation is not uncommon, and involves more than just individuals that posses extra copies of the same gene (e.g., Frazer et al. 2009; Zhang et al. 2009; Conrad et al. 2010). More dramatically, CNV can involve presence/absence differences of *entire genes*, as has been seen between maize lines by Fu and Dooner (2002). Hence, the actual composition of unique genes can vary over individuals. As a result, the notion of a single **reference genome** for a species is now replaced with the notion of a **pan genome**, the aggregate of several sampled genomes to capture all of the distinct genes within a population.

5. Large scale variation. Finally, very large-scale structural variation (at least at the level of moderately sized chromosomal segments)—such as inversions, large duplications, or translocations—can be seen within a sample. Such features are usually over a scale that is not useful for mapping beyond the crude chromosomal assays discussed at the beginning of Chapter 17.

While all of the above classes of DNA variation have been employed for linkage mapping, relationship estimation is almost entirely based on SNPs and SSRs. When possible, ***SNPs are the markers of choice in most settings***, because of their genome-wide abundance, their ease of rapid scoring in very large numbers using modern approaches (e.g., DNA chips and various next-generation sequencing technologies), and their stability.

Legacy Marker Systems

While whole-genome sequences contain all of the available relationship and linkage information in a sample, much of this variation is entirely redundant due the complete correlation of tightly-linked markers (LD r^2 values of one; Chapter 5). Hence, for many quantitative genetic applications, scoring only a subset of genome variation is sufficient. For example, with linkage-based mapping (such as a traditional line-cross QTL analysis;

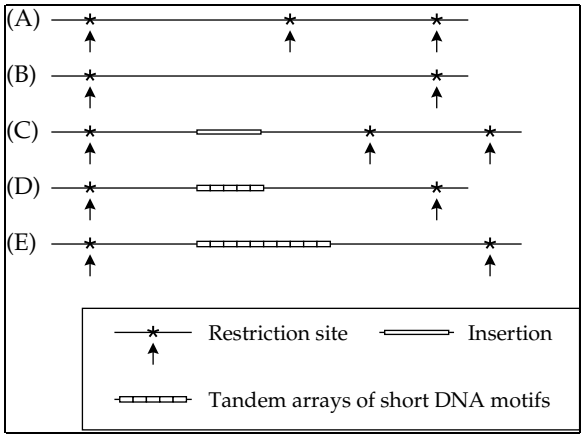

Figure 8.1 A variety of mutational events generate variation in restriction fragment lengths, scored by the distances between adjacent restriction sites. Single (or multiple) base-pair changes can create or destroy restriction sites, (A) vs. (B). Insertions of mobile genetic elements can create dramatic size differences (C). DNA sequences often exist as tandem arrays of short repeated sequences. Unequal crossing over and/or replication slippage creates variation in the size of such arrays, (D) vs. (E).

Chapter 18), the expected block size of nonrecombining regions is large enough that only a modest number of markers are needed to cover the genome. Thus, while we have the ability to score massive numbers of SNPs and STRs, this scale is often not required, and a number of other approaches for scoring DNA variation were used in early marker-based studies. We refer to these older approaches as **legacy marker systems**, as they are occasionally (albeit it diminishingly) still used today by researchers due to the legacy impact of having an existing set of markers on hand and ready to deploy.

One of the simplest approaches is to **digest** DNA with a variety of **restriction enzymes**, each of which cuts the DNA at a specific sequence or **restriction site** (commonly four to six bases). When the digested DNA is run on a gel under an electric current, the fragments separate out according to size. As Figure 8.1 shows, a variety of mutational events can generate fragment length variation. If we attempted to score the entire genome for fragment lengths, the result would be a complete (and uninformative) smear on the gel. Instead, individual bands are isolated from this smear by using labeled **DNA probes** that have base-pair complementarity to particular regions of the genome (often chosen at random). This approach is the basis for assays of **restriction fragment length polymorphisms** or **RFLPs** (Figure 8.2). Each RFLP probe generally scores a single marker locus with alleles being codominant, as heterozygotes and homozygotes can be distinguished. The number of detectable RFLPs is impressive (Botstein et al. 1980; Doris-Keller et al. 1987; Beckmann and Soller 1983, 1986a, 1986b; Soller and Beckmann 1988).

A particularly interesting RFLP approach involves using tissue-specific cDNA clones as the probes. **cDNAs** are generated from the mRNAs of genes being expressed in that tissue, and probes that detect cDNAs are often called **expressed sequence tags** (**ETS**). This procedure can allow for enrichment of genes of potential interest in the marker pool, enabling observed differences to be treated as potential candidate loci. For example, Kinzer at al. (1990) used 19 ripening-specific cDNA clones from the tomato (*Lycopersicon esculentum*) to detect polymorphisms in these loci between *L. esculentum* and a wild relative *L. pennellii*. One limitation with using ETS is that much of quantitative-genetic variation may be regulatory in nature, with causative sites residing outside of (and potentially very far away from) coding regions (Chapter 21). This is less of an issue with linkage mapping due to the large size of the LD blocks generated by a cross, but may be a serious issue in settings where LD blocks are expected to be rather small (such as in a population-level GWAS).

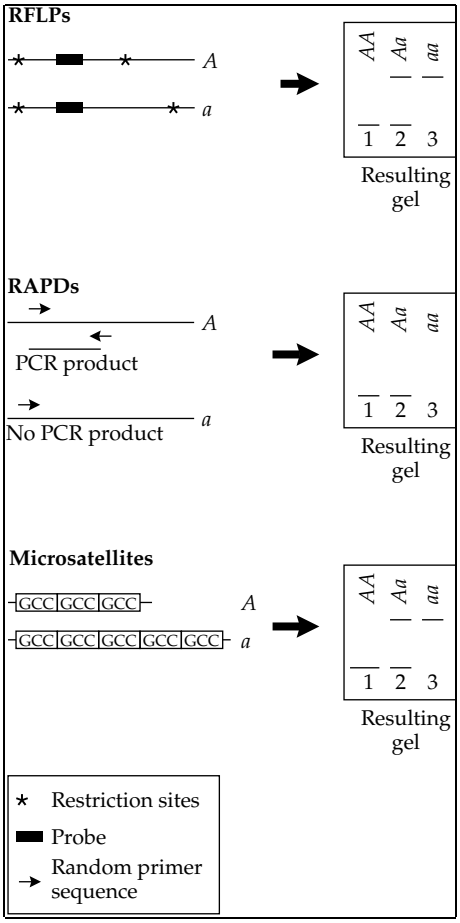

Figure 8.2 **Top**: RFLPs, restriction fragment length polymorphisms, variable lengths of restriction fragments in a particular region of the genome, are revealed by use of a probe for that region. This marker system is codominant (all marker genotypes can be distinguished), as an RFLP heterozygote shows fragments of both lengths. **Middle**: RAPDs, randomly amplified polymorphic DNAs, are revealed by the use of a random primer sequence. A successful PCR reaction requires sequences complementary to the primer in opposite orientation at sufficiently close distance. Running these products out on a gel, both *AA* and *Aa* exhibit a fragment and the site is scored as present, while the fragment does not amplify in *aa* individuals and the site is scored as absent. Hence, RAPDs are dominant markers. **Bottom**: Microsatellite DNAs exhibit variation in the array lengths of short sequences of tandemly repeated DNAs. Fragment length is scored, so that these markers are also codominant.

A rather different molecular marker assay uses short **primers** for DNA replication via the **polymerase chain reaction** (**PCR**) to delimit fragment sizes. A region flanked (in opposite orientation) by primer binding sequences that are sufficiently close together allows the PCR reaction to replicate this region, generating an amplified fragment. If primer binding sites are missing or are too far apart, the PCR reaction fails and no fragments are generated for that region. This procedure is the basis for **randomly amplified polymorphic DNAs** or **RAPDs** (Williams et al. 1990), sequence polymorphisms detected by using random short sequences as primers (Figure 8.2). RAPDs have an advantage over RFLPs in that a single probe (here, a particular random primer) can reveal several loci at once, each corresponding to different regions of the genome with the appropriate primer sites. However, by their very nature, RAPDs are generally dominant, as they indicate presence/absence of a particular site, so marker genotypes can be ambiguous (see Figure 8.2).

These basic components of generating markers—scoring the length of various DNA fragments and either amplifying or detecting a specific sequence using DNA hybridization generated by the base-pair complementarity—have been used to generate a number of other legacy marker systems (reviewed in Rafalski and Tingey 1993; Semagn et al. 2006; Grover and Sharma 2016). For example, several mapping studies have used mobile genetic elements (such as retroviruses) as markers (e.g., Rise et al. 1991; Nuzhdin et al. 1993; Keightley and Bulfield 1993; Ebert et al. 1993; Long et al. 1995). Due to the high rates of movement of some transposable elements, individuals often differ in the presence or absence of elements at particular sites. Scoring their presence/absence by simple hybridization yields dominant markers.

Rapid Scoring of a Large Number of SNPs

As mentioned, the current markers of choice are generally SNPs. One common approach for simultaneously, and inexpensively, scoring a very large number of SNPs is to use **chip technology**, which refers to a number a microarray approaches for scoring tens-of-thousands to millions of individual SNPs via hybridization technology. Once such a chip is constructed, the same panel of SNPs can be easily scored repetitively over large numbers of individuals. The main limitation is in the set of SNPs chosen to construct the chip, which tend to be biased towards SNPs with MAFs ≥ 0.05 (i.e., common SNPs). A second approach is **genotyping by sequencing** (**GBS**), a hybridization and amplification technology that can score a large number of random SNPs in an inexpensive fashion (Elshire et al. 2011; Poland and Rife 2012). Likewise, a variety of next-generation whole-genome sequencing approaches are available (e.g., Day-Williams and Zeggini 2010). Inevitably, quicker and cheaper new technologies will have appeared by the time this book is in print.

Marker information from SNP chip arrays and whole genome sequences have some fundamental differences (Browning and Browning 2013). The former scores up to a few millions SNPs, with a bias toward common SNPs (MAFs ≥ 0.01 to 0.05), and relatively modest error calling rates. The latter scores hundreds of millions of SNPs (in humans), the majority of which are rare (MAF ≤ 0.005), and with a much higher error rate, which is especially problematic in that a detected rare variant could be a true SNP or a sequencing error. Common alleles tend to be older mutations, while rare alleles tend to be newer mutations (WL Chapter 2). Hence, the two assay systems, arrays versus sequencing, may be interrogating mutations of different age classes, with common alleles shared by more distant relatives, and rare alleles shared by more recent relationships.

MARKER-BASED ASSIGNMENT OF CATEGORICAL RELATIONSHIPS

Methods estimating relationships from marker data can be grouped into two approaches: those that are *hypothesis-driven* (e.g., tests for paternity from a pool of candidate males, or determining whether family members are half- or full-sibs) and those that make *no prior assumptions* about relatedness. These two approaches can be restated as a focus on **categorical relationships** (***assigning*** pairs of individuals into *discrete classes* such as parent-offspring, or full- or half-sibs) versus continuous measures of **relatedness** (such as ***estimating*** the coefficient of coancestry, Θ). Relationship assignment is typically based on maximum likelihood methods (Appendix 4) and places individuals into a small number of distinct relationships, such as half sib, full sib, or unrelated (Thompson 1975, 1986a; Thompson and Meagher 1987; Boehnke and Cox 1997; Painter 1997; Broman and Weber 1998; Epstein et al. 2000; Sieberts et al. 2002; Wang 2004; Wang and Santure 2009). These approaches are commonly used in human genetics and forensics. They are also used in the quantitative genetics of natural populations when family units are observed, but with uncertainty about paternity, and the related issue of whether sibs are half or full. This approach is typically used to verify connections or fill in missing links in a suspect or incomplete pedigree (Blouin 2003; Pemberton 2008; Huisman 2017). We first consider these assignment methods before turning our attention to estimation of relatedness in the remaining sections.

Paternity Assignment

In the study of natural populations (including humans), one can often assign offspring to a specific mother, forming a natural family unit. Examples include species with extensive maternal care (such as birds, mammals, and some invertebrates) and seeds recovered from an individual plant or tree. When more than one offspring occurs in a family, they could have a single father (full sibs) or multiple fathers (half sibs). A related issue, when applicable, is whether offspring are selfed or outcrossed. A variety of approaches have been proposed for paternity testing, many for the challenging situation where family units may not be obvious. Our focus here is on likelihood-based methods, while variants and other strategies are reviewed by Blouin (2003), Jones and Ardren (2003), Jones et al. (2010), Walling et al. (2010), and Harrison et al. (2013).

We use markers to assess candidate fathers by a two-step process. First, the marker data may **exclude** a candidate. For example, if the mother is M_2M_2 and her offspring is M_2M_3, then any candidate father who does not contain an M_3 allele is immediately excluded (short of a genotyping error). One common approach is to simply use exclusion alone, especially if all but one of the candidates is so excluded (e.g., Dodds et al. 1996; Wang 2007). Second, under **failure to exclude**, we can assign a likelihood ratio of being the father, as opposed to being unrelated to the offspring. We consider four settings for paternity analysis. In increasing order of complexity, these are: (i) known mother and single candidate father, (ii) unknown mother and single candidate father, (iii) known mother and multiple candidate fathers, and (iv) unknown mother and multiple candidate fathers.

We follow Meagher (1986) and Marshall et al. (1998), whose work builds on a large previous literature (see Meagher and Thompson 1986, 1987; Thompson 1986b; Thompson and Meagher 1987; Pena and Chakraborty 1994, and references there in), and uses likelihood ratios (Appendix 4). The basic idea is to contrast the probability (likelihood, L) of the data (D) under two alternate settings—H_1 (the candidate is the father) and H_2 (he is not)—considering the ratio

$$LR(H_1, H_2 \mid D) = \frac{L(D \mid H_1)}{L(D \mid H_2)} \tag{8.1}$$

This ratio quantifies how much more probable the offspring marker genotype is given that the candidate is the father, as opposed to the candidate being unrelated to the offspring. As stressed by Thompson (1986b), this ratio is "*not* a comparison of individuals (fathers and non-fathers) but to two genealogical hypotheses ('father' and 'unrelated') about a given man."

We build the likelihoods for Equation 8.1 as follows. Let the data $D = (g_m, g_o, g_f)$ denote, respectively, the single locus marker genotypes for the **trio** of mother, offspring, and candidate father (a *pair* of relatives, such parent and offspring, is called a **dyad**). Recalling from conditional probability (Chapter 3) that $\Pr(x, y) = \Pr(x \mid y) \Pr(y)$, we have

$$L(g_m, g_o, g_f \mid H_1) = T(g_o \mid g_m, g_f)\, P(g_m)\, P(g_f) \tag{8.2a}$$

$$L(g_m, g_o \mid H_2) = T(g_o \mid g_m)\, P(g_m)\, P(g_f) \tag{8.2b}$$

Note that we have assumed random mating, so that $P(g_m, g_f) = P(g_m)\, P(g_f)$. The key element of Equation 8.2 is the **transmission probability**, T, of obtaining the offspring genotype g_o given the genotype of one (mother, g_m, H_2) or both (mother and candidate father, g_f, H_1) parents. This is the trio transmission probability $T(g_o \mid g_m, g_f)$ under H_1 and the dyad transmission probability $T(g_o \mid g_m)$ under H_2.

To see how these transmission probabilities are computed, consider a biallelic SNP with alleles 0 and 1. For such a marker, homozygotes are denoted by 00 and 11, and heterozygotes by 10. If the mother is 00 and her offspring is 10, then the transmission probability associated with a 11 father is Pr(1 from father)·Pr(0 from mother) + Pr(0 from father)· Pr(1 from mother) = 1·1 + 0·0 = 1. In contrast, the associated transmission probability for a random father becomes p·1 + (1-p)·0 = p, where p denotes the frequency of the 1 allele. Hence,

$$T(g_o = 10 \mid g_m = 00, g_f = 11) = 1$$

$$T(g_o = 10 \mid g_m = 00) = p$$

Note that this requires accurate reference population allele frequencies, p. Marshall et al. (1998) presents tables of these transmission probabilities for all possible combinations of marker genotypes for the trio, $T(g_o \mid g_m, g_f)$, and for a mother and her offspring, $T(g_o \mid g_m)$.

Substituting Equations 8.2a and 8.2b into Equation 8.1,

$$LR(H_1, H_2 \mid D) = \frac{T(g_o \mid g_m, g_f)\, P(g_m)\, P(g_f)}{T(g_o \mid g_m)\, P(g_m)\, P(g_f)} = \frac{T(g_o \mid g_m, g_f)}{T(g_o \mid g_m)} \tag{8.3a}$$

which yields a value of $1/p$ for our trio data of $g_o = 10$, $g_m = 00$ and $g_f = 11$. Assuming loci assort independently and are in linkage equilibrium, the total likelihood ratio is simply the product of Equation 8.3a over all markers. This resulting product is often called the **paternity index**, **PI** (Pena and Chakraborty 1994). A common convention (e.g., Slate et al. 2000) is to take an LR value in excess of 20 as strong evidence for paternity (the candidate is 20 times more likely to generate the offspring if he is the father as opposed to being unrelated). This is often stated as a LOD score (natural log of the LR; Chapter 18) in excess of 3, which follows by noting that that $\ln(20) \simeq 3$. As a technical aside, the more sophisticated reader might wonder why we simply do not perform a standard LR test, using the large-sample approximation that Equation 8.3a asymptotically follows a χ^2_1 (Appendix 4). The problem is that the null hypothesis (the probability that a specified state is one) lies on the boundary of the parameter space, which compromises the standard χ^2 approximation (Appendix 4).

Using similar logic to that leading to Equation 8.2, when the mother is unknown, the likelihood ratio becomes

$$LR = \frac{T(g_o \mid g_f)\, P(g_f)}{P(g_o)\, P(g_f)} = \frac{T(g_o \mid g_f)}{P(g_o)} \tag{8.3b}$$

Consider the example from above, where the offspring is 10 and the father is 00. Here $P(g_o) = 2p(1-p)$ and $T(g_o \mid g_f)$ = Pr(1 from father)·Pr(0 from mother) + Pr(0 from father)·Pr(1 from mother) $= 1 \cdot (1-p) + 0 \cdot p = 1-p$, giving LR = $(1-p)/[2p(1-p)] = 1/(2p)$. As these examples illustrate, the known mother setting is more powerful than the unknown mother setting. They also show power is a function of marker allele frequencies, with the signal for IBD increasing as the allele becomes rarer. Hence, ***sharing of rare alleles is a strong signal for IBD sharing***.

While our examples have used SNPs for ease of exposition, paternity testing is generally done using SSRs, as the latter are highly polymorphic and consist a constellation of alleles of roughly equal frequency. Meagher (1986) and Meagher and Thompson (1986) noted that the power of a marker locus increases with its number of alleles and the evenenss of its allele frequencies (higher power when the frequencies are more equal). Hence, a typical SSR has more power than a typical SNP. Indeed, Städele and Vigilant (2016) estimated that one SSR has roughly the power of 6 well-choosen SNPs.

Example 8.1 Suppose we have an offspring, its mother, and two potential candidates for the father (male 1 and male 2), which have all been scored at four biallelic SNPs. The resulting genotypes are as follows:

Marker locus	1	2	3	4
Offspring (g_o)	00	10	11	10
Mother (g_m)	00	11	11	10
Male 1 (g_{f1})	11	10	11	10
Male 2 (g_{f2})	00	00	11	10
$T(g_o \mid g_m, g_{f1})$	0	0.5	1	0.5
$T(g_o \mid g_m, g_{f2})$	1	1	1	0.5

The transmission probabilities, *T*, follow using the logic above. For example, consider marker locus 4 and male 1. The parental genotypes are 10 and 10 and the offspring is also 10. This can

occur via two paths: 1 from mom and 0 from dad, each with probability 0.5, or 1 from dad and 0 from mom, giving $T = (0.5 \cdot 0.5) + (0.5 \cdot 0.5) = 0.5$. Note (short of a genotyping error) that marker locus 1 excludes male 1, as this male has transmission probability of zero.

To compute the LRs, we need the allele frequencies in the reference population, which then yields the components needed to apply Equation 8.3a and 8.3b as

Marker locus	1	2	3	4
p (Freq allele 1)	0.1	0.2	0.5	0.3
$T(g_o \mid g_{f2})$	0.9	0.2	0.5	0.5
$P(g_o)$	0.81	0.32	0.25	0.42
$T(g_o \mid g_m)$	0.9	0.8	0.5	0.5

Applying Equation 8.3a, the LR values for male 2 for the four markers are, respectively, $1/0.9 = 1.11$, $1/0.8 = 1.25$, $1/0.5 = 2$, and $0.5/0.5 = 1$. Assuming linkage equilibrium, multiplying these together give the final LR value as 2.78. Now suppose that the mother is unknown. Equation 8.3b gives the LR values for the four markers in male 2 as, respectively, $0.9/0.81 = 1.1$, $0.2/0.32 = 0.63$, $0.5/0.25 = 2$, and $0.5/0.42 = 1.2$, giving a total LR of 1.65.

If we are willing to assign a prior probability, π_o, that the candidate is the father, we can use Bayes' theorem (Equation 3.3b) to update this probability given the marker data (D). Following Li and Chakavarti (1985) and Thompson (1986b),

$$\Pr(\text{Paternity} \mid D) = \frac{\pi_o \Pr(D \mid H_1)}{\Pr(D)} = \frac{\pi_o \Pr(D \mid H_1)}{\pi_o \Pr(D \mid H_1) + (1-\pi_o)\Pr(D \mid H_2)} = \frac{\pi_o}{\pi_o + (1-\pi_o)/LR} \tag{8.4}$$

where LR is given by Equation 8.3a (known mother) or Equation 8.3b (unknown mother). In the simplest setting, if one has k potential fathers, we could take $\pi_o = 1/k$. More sophisticated approaches can be used that take into account additional information such time spent with the mother and other behavioral observations (e.g., Neff et al. 2001; Hadfield et al. 2006).

The more general paternity problem is when one has a number of candidate fathers. This commonly arises in natural populations (especially in plants) and has been examined by a number of authors (e.g., Meagher 1986; Meagher and Thompson 1986; Marshall et al. 1998). The first step is excluding candidates to narrow the pool of potential fathers. As mentioned, genotyping errors can result in false exclusions, as can the presence of **null alleles** (denoting such an allele by Ø, 0Ø and 1Ø are erroneously scored as 00 and 11, respectively). Genotype calling errors and null alleles can be built into the likelihood function to accommodate these concerns (e.g., Marshall et al. 1998; Sieberts et al. 2002; Wang 2004; Kalinowski et al. 2006, 2007; Jones and Wang 2016). The next step is to *rank* the nonexcluded males, by computing the LR for each candidate, with the male with the largest LR value taken as the father (the MLE, or maximum likelihood estimate; Appendix 4). If the mother is unknown, we can follow the same procedure, but now using Equation 8.3b to compute the LR.

While such an approach provides the ML solution for the most likely father among the candidate set, due to sampling, the actual father might not have been included. There is also the problem in deciding if the *most likely* male is *significantly better* than the next candidate (Meagher 1986; Marshall et al. 1998; Slate et al. 2000). Letting LR_1 and LR_2 denote the largest and second largest values, Marshall et al. suggested that the statistic $\Delta = \ln(LR_1) - \ln(LR_2)$ be used. In an attempt to account for multiple comparisons (Appendix 6), they evaluate the significance of Δ using simulations based on the number of candidate males and the proportion of males that are sampled. Alternative approaches for this multiple comparison problem have been suggested by Devlin et al. (1988, 1989), Smouse and Meagher (1994), Smouse et al. (1999), Neff et al. (2000a, 2000b), Nielsen et al. (2001), Geber et al. (2003), and Christie (2010).

Note that the above LR assignment approach assumes a male is either the father or is completely unrelated. In reality, *relatives* of a candidate and/or the offspring (in addition to its mother) may be included in the sample. This is not much of a problem for distant relatives, but complications can arises when close relatives are included in the candidate set. It is especially problematic when sibs of the true father, or undetected sibs of the offspring, are considered as potential fathers. For example, full sibs of an offspring can gave *higher* LR values than the true father, a complication that Olsen et al. (2001) referred to as **aunt and uncle** (**avuncular**) **effects**. When the mother is known, these concerns only arise with full sibs, but when the mother is unknown, half sibs also cause complications (Thompson and Meagher 1987; Marshall et al. 1998; Olsen et al. 2001; Ford and Williamson 2010). The simplest solution is to estimate Θ directly for each pair (see below), rather than trying to assign individuals into a few discrete relationship categories which may not have sufficiently granularity given the genealogy of the population.

Finally, the most challenging situation arises when *both* parents of an offspring are unknown. Meagher and Thompson (1986) suggested that one first uses Equation 8.3b to assess paternity and then also use it to assess maternity (using g_f is place of g_m). Among the candidate set of mothers and fathers that this procedure generates, one then tests all pairwise combinations of these high-value (large LR) parents, and the joint pair with the highest LR value is taken as the MLE of the parents. Meagher and Thompson showed this two-step approach is appropriate, as single-parent likelihoods are highly correlated with the joint-parent likelihoods.

Sibship Assignment and Joint Family Likelihoods

There are two rather different settings for sibship assignment. In the simple case, one has a natural family unit (typically a mother and her offspring) and the question is whether these offspring are all full sibs or a collection of full and half sibs (and, if the latter, the secondary question of how many distinct fathers were involved). In the more complex case, one has sampled a natural population of offspring with no obvious mother (or father) and the question is how many familial units are present, and then clustering individuals into collections of full sibs potentially nested within larger half-sib families.

While exclusion is possible with marker data from a parent-offspring dyad, such is *not* the case with a putative sib dyad. Here, exclusion would be ruling out the pair being half-sibs if the mother known; or ruling out being either half or full sibs when neither parent is prescribed. Suppose a candidate pair from the sample have genotypes of M_1M_2 and M_3M_4. Such a dyad could be half sibs, full sibs (the parents were M_1M_3 and M_2M_4), or unrelated. While simple exclusion is not possible from sib dyad data alone, we can compute a LR for these different categories (given marker allele frequency estimates). As one adds more individuals into the comparison, exclusion is possible, and especially if more putative sibs in the family are considered. This follows by noting the probability that both possible alleles from a parent are present in the genotypes of n offspring is $1-(1/2)^{n-1}$ (Wang 2007; Wang and Santure 2009).

As this discussion suggests, using multiple-individual (**joint**) likelihoods (units of multiple putative sibs) has significantly more power that pairwise (dyad) LR comparisons. Variants of this approach have been used by a variety of workers (Painter 1997; Almudevar and Field 1999; Thomas and Hill 2000; Thomas et al. 2000; Emery et al. 2001; Smith et al. 2001; Sieberts et al. 2002; Almudeva 2003; Bulter et al. 2004; Wang 2004, 2012; Fernández and Toro 2006; Wang and Santure 2009; Jones and Wang 2010; Almudevar and Anderson 2012; Huisman 2017). The basic logic of such a **full likelihood** approach follows by expanding the transmission probabilities used in simple paternity analysis. The complication is that one must run over all combinations of the possible transmission probabilities for a set of individuals under various assumption about their relatedness. This is computationally demanding, especially in the more complex setting of no obvious family units. Further, this process must be computed separately for each marker (with the resulting LRs multiplied together). These calculations can be approximated using MCMC-based approaches (such

as simulated annealing; Appendix 8), or by cleverly combining information from pairwise likelihoods (Wang 2012).

MARKER-BASED ESTIMATES OF RELATEDNESS: BACKGROUND

Finally, we turn from *assigning* pairs of individuals into *discrete relationship classes* to the *estimation* of *continuous measures* of *relatedness*. The possible single-locus IBD states between a pair of individuals are completely described by Jacquard's (1974) nine condensed coefficients of identity, Δ_i (Figure 7.2). These collapse to just three IBD states in the absence of inbreeding ($\Delta_7, \Delta_8, \Delta_9$; Chapter 7), which are denoted in the human genetics literature by the vector $\mathbf{k} = (k_0, k_1, k_2)$, corresponding to $(\Delta_9, \Delta_8, \Delta_7)$. Here k_i is the probability that the pair shares i IBD alleles, with $k_2 = 1$ corresponding to monozygotic twins, $k_1 = 1$ to parent-offspring, and $k_0 = 1$ to unrelated individuals. Note that some authors, following Cotterman (1940), denote Δ_8 as $2k_1$, and the k_i are often referred to as **Cotterman coefficients.** The most general setting would be to estimate the likelihoods (or, in a Bayesian setting, the posterior probabilities; Appendix 7) all the Δ_i coefficients for a specific pair (e.g., Example A4.7). However, in the vast majority of cases, our interest is in summary statistics of these identities, namely the coefficient of coancestry (Θ_{xy}; Equation 7.2), an individual's inbreeding coefficient (f_x), and, more infrequently, the coefficient of fraternity (Δ_{xy}).

The most common reason for estimating relationships is to obtain variance components (we treat genomic selection/prediction as an extension of variance estimation; Chapters 31 and 32). Recall that $2\Theta_{xy}$ and Δ_{xy} are, respectively, the weights on the additive and dominance variances in the phenotypic covariance between x and y (Chapter 7). More generally, the weight on the k-level additive and ℓ-level dominance variance component, $\sigma^2(A^k D^\ell)$, is $(2\Theta_{xy})^k \Delta_{xy}^\ell$ (Equation 7.12). We use these weights to translate phenotypic covariances into variance estimates (Chapters 22–29 and 32). When there is extensive inbreeding, this basic logic still applies, with the added complication that additional variance components arise. This is examined in detail in Chapter 11 of WL. A second important use of relationship estimates is to ***control for shared ancestry***, such as in a GWAS (Example 9.13; Chapter 20), where the goal is contrast the effects of a feature (such as a particular SNP) once the impact of cryptic relatedness is removed. A final use is in estimating the level to which a target individual is inbred (estimating its inbreeding coefficient, f). Chapter 12 examines the impact of inbreeding in detail.

Our discussion will first consider a few key conceptual issues, such as the choice of a base population in defining IBD status, IBD probability versus IBD correlational definitions of relatedness, and variance in realized relationships. We then consider a variety of approaches that have been proposed for marker-based estimates of relatedness. While the logic for most methods is relatively independent of the types of markers being used, they often reflect the historical transition from assuming only a few modestly informative markers (alloyzmes, which often contain multiple alleles at modest frequencies) to a few very informative markers (SSRs; e.g., Queller et al. 1993; Blouin et al. 1996) and, finally to a large to huge number of relatively low-information markers (SNPs). There is a very rich (and growing) literature on marker-based relationship estimation. Reviews and commentaries can be found in Ritland (2000), van de Casteele et al. (2001), Blouin (2003), Milligan (2003), Garant and Kruuk (2005), Csilléry et al. (2006), Oliehoek et al. (2006), Weir et al. (2006), Frentiu et al. (2008), Yang et al. (2010, 2011a), Sillanpää (2011), Toro et al. (2011); Gay et al. (2013), Bérénos et al. (2014), Jensen et al. (2014), Wang (2014, 2016), Conomos et al. (2016), Ackerman et al. (2017), Wang et al. (2017), Weir and Goudet (2017), and Goudet et al. (2018). Huang et al. (2014) provide extensions to polyploids. A masterful review of many of the issues considered here in a population genetics framework is given by Thompson (2013), while a different perspective on these issues is offered by Speed and Balding (2015).

IBD, Base Populations, and the Coalescent

A central organizing concept for treating genetic drift is the **coalescent** of a sample of alleles

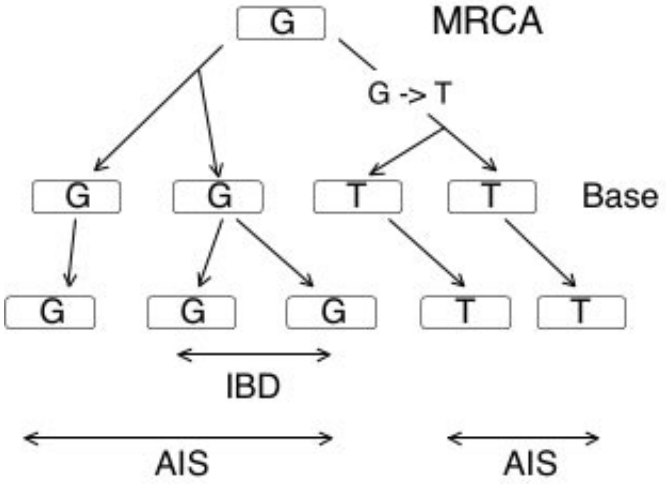

Figure 8.3 The genealogy, or **coalescent structure**, of a sample of five alleles at a single SNP. These all trace back to a single sequence in the most recent common ancestor (MRCA). Note that a mutation (from the **ancestral state** G to the **derived state** T) occurred, so that two of the sampled alleles differ in state from the MRCA. From a genealogical standpoint, all of the G alleles trace back to the MRCA (and hence are all IBD using that individual as the reference). Likewise, both T alleles trace back to a common ancestor (more recent than the MCRA), where the mutation occurred. At some point far more recent than the MRCA, a base population is designated. This is usually either taken as the present generation or no more than a modest number of generations back from the present. Conversely, the MRCA is on the order of $2N_e$ generations (N_e being the effective population size; WL Chapters 2 and 3). In the base population, all AIS alleles are assumed to be unrelated (not IBD). By moving the base population forward or backward in time, we change the IBD status of AIS alleles, even though the underlying coalescent structure remains unchanged.

(variants at a single SNP), namely their genealogy (the *pedigree* of the sampled alleles; Figure 8.3). Under coalescent theory, all existing alleles trace back to some **most recent common ancestor** (**MRCA**) through a series of coalescent events (reviewed in detail in Chapter 2 of WL). In particular, suppose that k sequences are sampled. At some point in the recent past, two of these sequences had a common ancestor, so that the k genealogies in the sample now coalesce to $k-1$ genealogies. The process continues until the last two remaining genealogies coalesce at the MRCA (Figure 8.3). Under the coalescent framework, all of the alleles are "IBD" to the MRCA, as coalescent theory is *only* concerned with the *genealogy*, and *not* the *allelic states* of its descendants, namely, the contrast between *relatedness* versus *identity*. Mutation is overlaid on this genealogy to generate distinct AIS allelic classes (Figure 8.3).

The strictly genealogical coalescent view of IBD is at odds with the usual notion of IBD status, which further assumes no mutation since the common ancestor, so that IBD alleles *must* also be AIS. As noted by Powell et al. (2010), this apparent conflict is simply a difference in reference points. In particular, we define some **base population** for IBD calculations (Figure 8.3). We assume that all individuals in the base are unrelated and not inbred (although we can modify either of these assumptions; Chapter 7). As shown in Figure 8.3, with respect to the specified base, only two of the sampled alleles are IBD, yet the sample contains multiple AIS alleles (two T and three G). With this same sample of five alleles, we can change the IBD status of these AIS alleles simply by moving the base forward or backward in time. Hence, while the MRCA is at a fixed (but unknown) generation, the generation chosen for the reference population is arbitrary, usually being specified for the convenience of the researcher. Very often, the base population is simply taken as the current population.

The reason we have focused on a single SNP in the above discussion is due to an important subtlety with the MRCA. It is ***defined for a specific site***, and hence changes over a sequence, in that when recombination occurs, ***the coalescent structures can differ between even closely linked sites***. Thus, there is usually not a common MRCA for all the sites over a gene, much less over a chromosome or genome. A related issue is the notion of an **IBD chromosomal segment**, namely, for the sampled sequences, an unrecombined region tracing back to a single copy in the reference population. In many populations, IBD segments are smaller than a typical gene-coding region. As we will see, segmental IBD is exploited for fine-mapping by GWAS (Chapter 20).

An important issue that we will not consider in detail here is the estimation of relatedness in highly structured populations, namely, when the larger population consists of a number of subpopulations (demes). This has been examined by Anderson and Weir (2007), Wang (2011b), Thornoton et al. (2012), Conomos et al. (2016), and Weir and Goudet (2017), and can be considered a special case of where one sets the reference population. Suppose we have a series of demes that have descended from a common ancestral population, and imagine a rare allele that is common in deme one. Setting the ancestral population as the base, individuals sharing this allele in deme one with be regarded as being closely related. However, if we set deme one as the base (and hence use deme one allele frequencies), the same set of individuals will be regarded, at best, as being weakly related (Anderson and Weir 2007).

There is one final important implication in setting the base population. As we have mentioned, a common use of relationship estimates is to obtain variance components. Estimates of the former (relatedness) vary with the base populations, which, in turn, influences estimates of the latter (variances), such that what we are estimating is the value of these variances in the base population. Hence, ***even when using the same data, as we change the assumed base population, we can change the variance estimates***.

Probabilistic versus Correlational Definitions of Relatedness

The inbreeding coefficient, f, underlies measures of relationship, both within an individual (f) and between individuals (Θ). For inbreeding, the comparison of IBD status is between the two alleles at a locus within an individual. For relationships, Θ_{xy} is based on choosing a random allele from the same locus in individuals x and y, with Θ_{xy} being the average value of f for the four combinations of allelic draws (Chapter 7). Equation 7.2b highlights this connection between inbreeding and coancestry: if z is the offspring of x and y, then $f_z = \Theta_{xy}$. Equation 7.3b provides a further connection in that $2\Theta_{zz} - 1 = f_z$.

Historically, there are two different definitions of f. While Wright (1922) originally defined f as the ***correlation between gametes***, the popular concept of f being the ***probability of IBD*** is due to Cotterman (1940) and Malécot (1948). Wright (1965), Powell et al. (2010), and Wang (2014) noted that while these concepts are identical in many cases (when the reference population is suitably ancestral to the current one), there are also settings where they differ. While IBD-probability estimates of f and Θ must lie between zero and one, the correlational definition can generate negative values, which in many situations are biologically realistic. A negative estimate of f simply implies that the focal individual is more heterozygous than expected by chance in the reference population. As noted by Wang (2014), one example is the F_1 hybrid of two inbred founder lines, which generates a negative value of f (due to the excess heterozygosity in the F_1), regardless of how the parental information is combined to form a reference population. Similarly, if the reference population contains a number of inbred individuals, then an outbred individual can a have a negative estimate of f (albeit typically very small), as it is likely to more heterozygous than a random member of the reference population. Similarly, negative estimates of Θ simply imply that two individuals are *more dissimilar* than two random individuals, again in the context of a reference population.

Powell et al. (2010) and Wang (2014) argued that the correlation-based definition of f (and, therefore, Θ) is the most appropriate choice for marker-based estimates that take the

current population (and hence the current allele frequencies) as the reference population. For known pedigrees, the obvious choice for a reference population is the collection of founders, but this raises issues of the rather arbitrary way in which founders are chosen. Typically, this is by convenience—known information—rather than from biological motivations. An example of the latter would be an isolated population with known founders (such as a captively bred population). More generally, it is convenient to simply take the current estimated allele frequencies as the reference population, which is assumed in most marker-based studies. Powell et al. (2010) showed the connection between f values associated with different base populations. Let f_1 and f_2 be the values based on two different assumed base populations, and let f_{st} be the correlation between a random gamete drawn from population one and a random gamete drawn from population two. Then

$$f_1 = \frac{f_2 - f_{st}}{1 - f_{st}} \tag{8.5}$$

Additional commentary on these issues is offered by Rousset (2002), Thompson (2013), and Speed and Balding (2015).

Finally, an important point was made by Powell et al. (2010), who argued that ***the objective of most marker-based IBD estimators is to predict the AIS status at the unobserved loci underlying the trait (or traits) of interest***. Obviously, if we knew these causative loci, we could score them directly. If not, we can try to predict their amount of AIS sharing using the marker data as proxies: individuals that share more marker loci IBD are likely to share more AIS causative alleles. A subtle caveat (which will be examined more fully in Chapters 20 and 21) is that if the distribution of marker allele frequencies (the **allele frequency spectrum**; LW Chapter 2) departs significantly from the distribution of causal allele frequencies, this lowers the quality of marker predictions. Such potential differences in frequency spectra are certainly not unexpected. First, there is often an ascertainment bias wherein common SNPs (minor allele frequency at least one to five percent) are chosen as markers, skewing the frequency distribution towards higher values. Second, it is thought that many (or most) SNP markers are selectively neutral (or nearly so), while the distribution of causative allele frequencies often reflects past selection (Chapter 21). This includes both direct artificial selection on the trait and prior natural selection. For example, alleles of large effect tend to be at low frequencies in natural populations, likely because they have been under negative natural selection in the past. IBD estimates from pedigree data (which estimate IBD independent of the underlying genetic architecture) are robust to this effect (Powell et al. 2010).

Realized versus Pedigree Kinship

The use of dense marker data highlights the important distinction, introduced in Chapter 7, between **pedigree kinship** and **realized kinship** (Wang et al. 2017). The former is the *expected* value of Θ calculated using a known pedigree (Chapter 7), while the latter is the *actual realization* of Θ for any particular pair of individuals, which can be estimated using dense marker data (e.g., Figure 7.5). This variation about the expected value is generated by **Mendelian sampling** (the segregation of the diploid genotype via meiosis into haploid gametes). Such variation occurs whenever at least two of the Δ_i probabilities are nonzero (which is always true except for clones and parent-offspring; Chapter 7). Dense SNP data can be used to capture this variation in relatedness, giving more accurate weights when using information from relatives (replacing expected values by their actual realizations). This is the basis of the genomic selection method known as **G-BLUP** (**genomic-BLUP**), wherein marker-estimated relatedness values are used in place of pedigree-generated values to improve the BLUP estimates (e.g., VanRaden 2007, 2008; Hayes et al. 2009; Chapters 31 and 32).

Equation 7.7b gives the expected variance in Θ generated by Mendelian sampling at a single locus, $\sigma^2(\Theta)$. With n independently segregating loci, the associated variance for the average of Θ over these loci is simply $\sigma^2(\Theta)/n$. Because chromosomes (generally)

assort independently, this result also applies when we sum chromosome-wide estimates. Consider the genome-wide value of Θ for a given pair of individuals (x, y) for a species with k chromosomes,

$$\widehat{\Theta}_{xy} = \sum_{i=1}^{k} w_i \, \widehat{\Theta}_{xy,i} \tag{8.6a}$$

where $\widehat{\Theta}_{xy,i}$ is the estimate for chromosome i, and w_i is the weight placed on this chromosome. Typically, if ℓ_i is the map length of chromosome i, and $\mathcal{L} = \sum^k \ell_i$ is the total genome map length, then $w_i = \ell_i/\mathcal{L}$. Chapter 17 discusses **map lengths—genetic distances** based on recombination, as opposed to **physical distances** based on base pairs—in detail. $\mathcal{L}$ is usually expressed in units of Morgans (the expected total number of crossovers) or centi-Morgans, cM, (100 cM = 1 Morgan.) The associated genome-wide variance becomes (Risch and Lange 1979)

$$\sigma^2(\widehat{\Theta}_{xy}) = \sum_{i=1}^{k} w_i^2 \, \sigma^2(\widehat{\Theta}_{xy,i}) \tag{8.6b}$$

The delicate issue is that, due to linkage, loci (or markers) on a chromosome are *not* independent. Hence, if we have n loci on a specific chromosome, the variance associated with that chromosome is *not* $\sigma^2(\Theta)/n$, as there are not n independent sampling events. A number of approximations have been proposed to accommodate this complication from linked, and hence nonindependent, sites (Franklin 1977; Risch and Lange 1979; Suarez et al. 1979; Stam 1980; Stam and Zeven 1981; Donnelly 1983; Hill 1993a, 1993b; Guo 1994, 1995, 1996; Visscher 1996, 2009; Visscher et al. 2006; Hill and Weir 2011, 2012; Thompson 2013). Most of this work is build around assuming the Haldane mapping function to model the recombination process (Example 8.3). As is discussed in detail in Chapter 17, this assumes a Poisson distribution for the number of crossovers within a region and no interference. Note that Equation 8.6b only refers to the **evolutionary variance**, the randomness associated with the Mendelian sampling process alone. There is an additional error, which we ignore here, generated by using a finite number of markers to estimate the Θ value for a given chromosome (or, more widely, genome).

One approach for obtaining the genome-wide variance is to use the single-locus variance (Equation 7.7b) in conjunction with some estimate, n_s, of the number of independently segregating chromosomal segments, taking the variance as $\sigma^2(\Theta)/n_s$. This approach is an approximation, attempting to model the recombinational behavior of a continuous chromosome by using a equivalent number of independently segregating segments (Donnelly 1983). Nonetheless, this is still a useful device for inquiring about some of the underlying factors influencing the variance in Θ. For relatives separated by a single round of meiosis (such as sib pairs), a simple metric for the number of segments is the **recombination** (or **segregation**) **index** (**RI**; Darlington 1939), which is the haploid chromosome number (k) plus the total number of crossovers per gamete (C), with RI = k + C. One can think of this as an approximation for the number of independent blocks of genetic material in a gamete.

As the number of generations back to the common ancestor increases, there is more opportunity for recombination, increasing n_s and thus decreasing the variance. This raises an important point. ***Sets of relatives with the same expected value of*** Θ (such as half sibs versus double first cousins) ***can have different expected standard deviations***. This occurs because although the single locus variance (Equation 7.7a) is the same in these cases, the expected number of recombinational events differs over the paths separating the compared relatives, resulting in differences in n_s values (Rasmuson 1993; Hill and Weir 2011). An equivalent way of thinking about this is that while the single gene IBD probabilities can be the same, the two-locus IBD probabilities differ (Thompson 1988, 2013). A related issue is that the chromosomal-specific variance decreases with the total amount of recombination on that chromosome. Visscher (2006) observed this in humans, and found that the predicted values for $\sigma^2(\widehat{\Theta}_{xy,i})$ based on the map length of the chromosome (Example 8.3) generally correlated well with their empirical values. More generally, the genome-wide variance in Θ

increases as chromosome number and/or total map length decreases, both of which have the effect of decreasing n_s.

Hill and Weir (2011) noted that both the skew and the CV increases as relatives become more distant (recall that Example 7.3 showed that the CV of Θ for half sibs was greater than for full sibs). As they succinctly stated: "Although the variance of actual relationships falls as individuals become more distant, its coefficient of variation rises, and so, exacerbated by the skewness, it becomes increasingly difficult to distinguish different pedigree relations from the actual fraction of the genome shared." Hence, ***the concept of discrete relationship classes quickly breaks down as relatives become more distant***, being replaced by continuous measures of relatedness that capture the realized values.

Example 8.2 A number of authors have considered the variance in the fraction of IBD genome sharing between a pair of full sibs (the common human setting). The IBD fraction is given by 2Θ (as Θ concerns drawing single alleles), and hence (Equation 3.10e) this variance is four times the variance of Θ. From Example 7.3, the single-locus variance for noninbred full sibs is $1/32$, giving $\sigma^2(2\Theta) = 1/8$, for a standard deviation of 0.354. Using data on the number of crossovers on each human chromosome, Suarez et al (1979) obtained a standard deviation for genome IBD of 0.056 (corresponding to $n_s = 40$) when allowing for recombination, and a value of 0.082 ($n_s = 19$) under the assumption of no recombination (chromosomal segregation only). Using an analytic model of the crossover process, Risch and Lange (1979) obtained a value of 0.04 ($n_s = 78$). Rasmuson (1993) used the estimated segregation index for humans (75 in males and 100 in females) to approximate the standard deviation as $0.354/\sqrt{100} = 0.0354$. Empirical estimates from Visscher et al. (2006) using 4,400 pairs of quasi-independent sibs gave a value of 0.036, with realized values of 2Θ ranging from 0.375 to 0.617. Their analytic results (Equation 8.7a) suggested that the effective number of segments per chromosome (n_s) ranged from a high value of 6 for the largest chromosome down to a value of 2 for the smallest, for a total number of 85 effective chromosome segments.

Example 8.3 To see the logic underlying approximations accounting for linkage, we develop expressions for full- and half-sibs variances for the fraction of IBD sharing, 2Θ. We assume no recombinational interference (independence of crossovers) and a Poisson distribution for the number of crossovers along a chromosome (the Haldane map function assumptions; Chapter 17). Consider loci i and j linked together on a chromosome. Recalling Example 7.3, for a single locus in a half-sib, $\sigma^2(2\Theta_i) = 4/64 = 1/16$. For two loci,

$$E[2\Theta_i 2\Theta_j] = [2(1 - c_{ij})^2 + 2c_{ij}^2]/16$$

where c_{ij} is the recombination fraction between i and j. Hence,

$$\sigma(2\Theta_i, 2\Theta_j) = E[2\Theta_i 2\Theta_j] - E[2\Theta_i]\,E[2\Theta_j] = \frac{(1 - 2c_{ij})^2}{16} = \frac{\exp(-4m_{ij})}{16}$$

where m_{ij} is the Haldane map distance (this follows from Equation 17.4). Summing over all n loci on a chromosome, the variance in the IBD fraction for that chromosome becomes

$$\sigma^2(2\Theta) = n^{-2}(1/16)\sum_i \sum_j \exp(-4m_{ij}) \simeq \frac{4\ell - 1 + \exp(-4\ell)}{128\,\ell^2} \tag{8.7a}$$

where ℓ is the total map length of the chromosome (in Morgans). Details for the last approximation (which follows for large n) can be found in Guo (1996) and Hill (1993a).

Suppose that are k chromosomes, the ith of which has length ℓ_i, for a total genome map length of $\mathcal{L} = \sum \ell_i$. Summing over Equation 8.7a yields the variance for the genome fraction of shared IBD alleles in half sibs as

$$\frac{1}{128\mathcal{L}^2}\left[4\mathcal{L} - k + \sum_{i=1}^{k} \exp(-4\ell_i)\right] \tag{8.7b}$$

which can be further approximated as $(4\mathcal{L} - k)/[128\mathcal{L}^2]$. Note that this shows the variance increases with decreasing values of k and/or $\mathcal{L}$. For human autosomes, $k = 22$ and $\mathcal{L} = 35$, giving an approximate variance of

$$(4 \cdot 35 - 22)/[128 \cdot 35^2] = 0.00075$$

for a standard deviation of 0.0274. Visscher (2009) noted that this approximation only differs at the fourth decimal place, 0.0272 versus 0.0274, when compared to the more exact value from Equation 8.7b. The latter can be computed by either using known values of ℓ_i, or by setting $\ell_i = 35/22$ (treating all chromosomes has having the same map length), as (in this case), both returned the same value (0.0272).

Similarly, using the same logic for full sibs yields

$$\frac{1}{64\mathcal{L}^2}\left[4\mathcal{L} - k + \sum_{i=1}^{k} \exp(-4\ell_i)\right] \tag{8.7c}$$

which can be further approximated as $(4\mathcal{L} - k)/[64\mathcal{L}^2]$, yielding a variance of 0.0015 and an associated standard deviation of 0.039. Nice summary tables of the variance for a wide variety of other relatives are given by Guo (1996) and Hill and Weir (2011). The latter also summarize expressions for the skewness.

Local versus Global Relatedness

The combination of Mendelian sampling (generating variation in realized relationships) coupled with the use of dense markers to detect such variation provides a very powerful tool. Consider a collection of full sibs. Because of this sampling, some will be more related, and hence (for a heritable trait) more phenotypically similar, that others. Visscher et al. (2006) used this within-sibship variation to estimate the heritability of human height. An even more granular analysis was proposed by Visscher et al. (2007). Just as Mendelian sampling can generate variation in the realized values of the **global relatedness** (the genome-wide average), it also generates *between-chromosomal differences* in the realized relationship *within* a pair of individuals (**localized relatedness**). Hence, we can estimate the heritability attributed to a *specific chromosome* by contrasting relatedness differences in a sample over just that chromosome. Because longer chromosomes have more recombination, and thus a smaller between-individual variance in fraction of shared IBD, smaller chromosomes tend to show more variance in relatedness, and therefore potentially more power to detect a chromosomal effect. A second implication for these between-chromosomal differences in variance is that smaller chromosomes in distant relatives are more likely to share no IBD segments than are larger chromosomes (Example 8.4).

Example 8.4. The connection between localized relatedness and variation in relatedness is through the notion of IBD chromosome segments. These are continuous stretches of DNA shared by the two focal individuals that have not undergone recombination since the common ancestor (**identity by recombination**). Following Donnelly (1983), Thompson (2013), and Speed and Balding (2015), consider two individuals that last shared a common ancestor g generations in the past. Further suppose there are k chromosomes, a total map length of $\mathcal{L}$, and a total genome size of M megabases (Mb). As relatives become more distant, the probability that they share *no* IBD segments increases. Thompson (2013) showed that one can approximate this by

$$\Pr(\Theta = 0) \simeq \exp\left[-\frac{(2g-1)\mathcal{L}}{2^{2g-1}}\right] \tag{8.8a}$$

Speed and Balding assumed human map lengths of 40.7 Morgans in females and 22.9 Morgans in males, taking $\mathcal{L}$ as their average, 31.8, and a value of $M = 2667$ Mb. For two individuals

that last shared a common ancestor five generations in the past,

$$\Pr(\Theta = 0) \simeq \exp\left[-\frac{(2 \cdot 5 - 1) \cdot 31.8}{2^{2 \cdot 5 - 1}}\right] = 0.67$$

Hence, the *majority* of such relatives will share *no* IBD segments. When the common ancestor is 3 generations back, Equation 8.8a yields 0.007, so that 99.3% of all such pairs will share *at least* one IBD segment.

A closely related issue is the number, n_s, of shared IBD segments. Speed and Balding approximated this as

$$E[n_S] \simeq \frac{k + 2\mathcal{L}g}{2^{2g-1}} \tag{8.8b}$$

Again taking $g = 5$, $E[n_S] = (22 + 2 \cdot 31.8 \cdot 5)/(2^9) = 0.66$, while for $g = 3$, $E[n_S] = 6.65$. These values imply that most chromosomes in both of these settings will share *no* IBD segments, generating considerable variance in IBD *between* the chromosomes of a pair of relatives. To further emphasize this point, consider the expected mean size (in megabases) of any such shared segment. Speed and Balding showed that the mean length, *conditional* on at least one shared region being present, is

$$\mu_S \simeq \frac{M}{k + 2\mathcal{L}g} \tag{8.8c}$$

Even though the majority of individuals with a common relative 5 generations in the past share no IBD segments, *when* they do, its average length is $2667/(22 + 2 \cdot 31.8 \cdot 5) = 7.8$ Mb. For $g = 3$, this conditional mean length is 12.5 Mb. This process, wherein fewer and fewer relatives share any IBD segments, but those that do have segments of at least modest size, is what generates the ever increasing skewness in IBD as relatives become more distant (Hill and Weir 2011). Many of the above results appear in the literature as function of the total number, m, of meioses that separate a pair from their common ancestor, with m replacing $2g$ in above expressions (Equation 8.8a–8.8c). Chapter 17 examines further examines the lengths of shared IBD regions.

MARKER-BASED ESTIMATES OF THE COEFFICIENT OF COANCESTRY, Θ

Finally, we turn to the estimation of relatedness statistics from marker data. Again, the issue is how to translate observed AIS sharing at the markers into estimates of IBD sharing at either a global (genome-wide) or local (chromosomal segment) level. This usually requires us to specify a reference population (but see Weir and Goudet 2017; Goudet et al. 2018), with the allele and genotype frequencies in the reference providing us with the expected amount of AIS sharing between random individuals. For a random-mating (i.e., unstructured) population, allele frequencies are sufficient to specify the expected AIS sharing. However, when there is structure in the base population or in the sample (such as inbreeding or when a substantial fraction of the sampled individuals are relatives), allele frequencies alone are not sufficient to adjust AIS status, and we will examine how to accommodate these concerns.

Our discussion largely frames estimators in terms of biallelic SNP markers, but most can easily be extended to multiallelic markers, such as SSRs. Most of these approaches use **single-point** (or **SNP-by-SNP**) **estimates**, which treat markers as being independent and then amalgamates these single-locus results into a final composite estimate. Note that single-point estimators do not require any mapping (i.e., positional) information on the markers. As marker density increases, nearby markers become increasingly correlated (linkage disequilibrium; Chapters 5 and 17). This is typically dealt with by **trimming** the number of markers to form a set of largely independent loci (e.g., Purcell et al. 2007; Huisman et al. 2016), but this loses information. More recent estimators exploit information provided by the pattern of linkage disequilibrium, either by weighting adjacent markers by their correlations (e.g., Day-Williams et al. 2011; Speed et al. 2012; Wang et al. 2017) or using the

IBD status of *chromosomal segments*, also known as **recombinational IBD** (Example 8.4), as opposed to using IBD *alleles* at individual loci (reviewed by Browning and Browning 2012).

Our focus will be on estimating the three most widely reported relatedness metrics: the coefficient of coancestry (Θ_{xy}; which has the fraction of IBD genome sharing, $2\Theta_{xy}$, as a special case), the inbreeding coefficient of an individual (f_x), and the coefficient of fraternity (Δ_{xy}). We consider these in turn, largely focusing on **method of moments** estimators, wherein one generates statistics whose expected value equals the parameter of interest. Maximum likelihood provides a more rigorous statistical framework, but can be computationally quite demanding. Appendix 4 examines ML estimates of continuous relationship metrics.

Single-point IBD Estimators

There are a number of proposed methods that translate data from a SNP into estimates of Θ. The basic approach is as follows. Consider two individuals, x and y. We denote the two alleles in x by a and b (which may be alike in state), and similarly in y by c and d. The **molecular similarity** (or **molecular similarity index**) at locus ℓ between x and y is defined by

$$S_{xy,\ell} = \frac{I_{ac} + I_{ad} + I_{bc} + I_{bd}}{4} \tag{8.9a}$$

where I_{ad} is an indicator function that equals one if a and d are AIS, and otherwise is zero. For biallelic loci (such as most SNPs), $S_{xy,\ell}$ takes on values of 0, 1/2, or 1. A value of zero occurs if x and y are different homozygotes, while $S_{xy,\ell} = 1$ if they are the same homozygote. All other biallelic locus combinations yield a value of 1/2. A value of 1/4 requires at least three distinct alleles, and values of 3/4 do not occur, because if the first three combinations are one, so is the last (Oliehoek al. 2006). Toro et al. (2002) referred to Equation 8.9a as **molecular coancestry**, as when AIS equals IBD, then $E[S_{xy,\ell}] = \Theta_{xy}$, which follows from the definition of the coefficient of coancestry (Chapter 7). The average over L marker loci yields an estimate of

$$\widehat{\Theta}_{xy} = \frac{1}{L}\sum_{\ell=1}^{L} S_{xy,\ell} \tag{8.9b}$$

The problem with this simple estimator is the equating of AIS with IBD, as unrelated individuals can share AIS alleles. One needs to assign a base (or reference) population and use the expected genotype frequencies in this base to provide a correction for AIS status among random individuals. Any such adjustment makes some key assumptions. Usually these are that one has a random sample from the true assumed reference population, and that individuals within the sample are largely unrelated and largely outbred. If the sample contains a sufficient number of inbred individuals and/or relatives, this can bias the correction for alleles that are AIS but not IBD. While this bias may be small, it can potentially be rather large, especially in small populations or small samples (Wang 2002; Goudet et al. 2018). A related, but subtle, issue is that the assumed base population *itself* (as opposed to the sample) may contain a significant fraction of relatives and inbred individuals (such as might be expected in captive-bred population). A random pair of individuals from this population has some reasonable chance of being related and/or inbred. IBD estimates in this setting are best thought of as ***differences in IBD status relative to that expected in random draw from the reference population*** (Goudet et al. 2018), and, as such, can take on negative values.

To proceed under the assumption of an appropriate sample from a random-mating reference population, let s_ℓ denote the probability that two randomly drawn alleles in the base are AIS. As shown by Lynch (1988c), the expected value for $S_{xy,\ell}$ is given by

$$E[\, S_{xy,\ell}\,] = \Theta_{xy} + (1 - \Theta_{xy})s_\ell \tag{8.10a}$$

namely, Θ if they are related (IBD, hence AIS), and s_ℓ if they are not related, the latter occurring with probability $1 - \Theta$ (not IBD). Here s_ℓ is the expected AIS sharing among a

randomly drawn pair of individuals from the reference population. Underestimating of the amount of AIS sharing between random individuals results in overestimation of relatedness, while overestimation of AIS results in underestimation of relatedness.

For a biallelic locus in Hardy-Weinberg,

$$s_\ell = p_\ell^2 + (1-p_\ell)^2 = 1 - 2p_\ell(1-p_\ell) \tag{8.10b}$$

where p_ℓ is the SNP allele frequency in the reference population, a value typically estimated from the sample. When this is done, Queller and Goodnight (1989) recommend first removing the pair under consideration from the sample, and then estimating p_ℓ using this reduced sample. While this is reasonable, the pair may share a rare allele, in the extreme only present in the pair, in which case the reduced sample estimate would be zero. If the sample size is modest to large, including the pair of interest when estimating p_ℓ should introduce little bias, provided the average coancestry in the sample is very close to zero.

Rearranging Equation 8.10a suggests a more general method of moments estimator (Lynch 1988)

$$\widehat{\Theta}_{xy,\ell} = \frac{S_{xy,\ell} - s_\ell}{1 - s_\ell} \tag{8.10c}$$

(note the similarly to Equation 8.5). If we weight all markers equally, then a composite estimate of Θ_{xy} based on L markers (either globally, with the markers spread across the genome, or locally, with the markers restricted to a chromosome or chromosomal segment) is given by

$$\widehat{\Theta}_{xy} = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{\Theta}_{xy,\ell} \tag{8.10d}$$

Lynch and Ritland (1999) suggested that equal weighting is not optimal (in terms of producing the smallest sampling variance), but rather each estimate should be weighted by its sampling error. Eding and Meuwissen (2001) obtained the sample variance of Equation 8.10c, assuming that s_ℓ is known without error, as

$$\sigma^2\left(\widehat{\Theta}_{xy,\ell}\right) = \frac{s_\ell + \Theta_{xy}(1-2s_\ell) - \Theta_{xy}^2(1-s_\ell)}{1-s_\ell} = w_\ell^{-1} \tag{8.10e}$$

Hence, a more optimal estimator is

$$\widehat{\Theta}_{xy} = \frac{1}{W}\sum_{\ell=1}^{L} w_\ell\, \widehat{\Theta}_{xy,\ell}, \quad \text{with} \quad W = \sum_{\ell=1}^{n} w_\ell \tag{8.10f}$$

Equation 8.10e shows that optimal weights are functions of the parameter we are attempting to estimate, but one could substitute an initial estimate and then iterate until convergence.

Negative estimates of Θ can arise when $S_{xy,\ell} < s_\ell$ over a large number of loci, implying that these individuals are *less related than expected by chance* (relative to random individuals drawn for the reference population). Assuming $s_\ell = 0$ eliminates this problem (the assumption behind Equation 8.9b), but also introduces bias (Speed and Balding 2015; Ackerman et al. 2017). Oliehoek al. (2006) obtained an adjusted value for s_ℓ to ensure that all the $\widehat{\Theta}_{xy}$ are nonnegative, but again this likely introduces some slight bias. Such attempts to force estimates to be nonnegative are largely based on the incorrect interpretation that Θ *must* be positive. As discussed above, negative estimates are both expected and biologically interpretable.

Alternatively, one can base an estimator on the *total number* of shared AIS alleles over L loci (Day-Williams et al. 2011). First, define

$$S_{xy} = \sum_{\ell=1}^{L} S_{xy,\ell} \tag{8.11a}$$

which is the total number of AIS alleles shared between x and y over all scored markers. This has expected value

$$E[\,S_{xy}\,] = \sum_{\ell=1}^{L} E[\,S_{xy,\ell}\,] = \sum_{\ell=1}^{L} [\Theta_{xy} + (1-\Theta_{xy})s_\ell] = L\Theta_{xy} + (1-\Theta_{xy})\sum_{\ell=1}^{L} s_\ell \qquad (8.11b)$$

Rearranging yields the **Day-Williams** estimator

$$\widehat{\Theta}_{xy,DW} = \frac{S_{xy} - \sum_{\ell=1}^{L} s_\ell}{L - \sum_{\ell=1}^{L} s_\ell} \qquad (8.11c)$$

The distinction between the estimators given by Equations 8.10d and 8.11c is that the former corrects for AIS on a *locus-by-locus* basis, while the latter corrects for the *summed total* of AIS over all loci. Equation 8.10b gives the values of s_ℓ under the assumption of an appropriate sample from a random-mating reference population.

When these assumptions about either the sample, or the base population itself, are incorrect, these AIS corrections can be biased. A simple, but robust, way around this complication is the **Weir-Goudet estimator** (2017). This can be expressed as either the total number (S) or proportion ($M = S/L$) of AIS alleles over L loci,

$$\widehat{\Theta}_{xy,WG} = \frac{S_{xy} - \overline{S}}{L - \overline{S}} = \frac{M_{xy} - \overline{M}}{1 - \overline{M}} \qquad (8.12a)$$

where n is the sample size and

$$\overline{S} = \frac{1}{n(n-1)} \sum_{j \neq i}^{n} \sum_{i=1}^{n} S_{ij} \qquad (8.12b)$$

is the average sharing over all pairs of individuals in the sample (with $\overline{M} = \overline{S}/L$). The Weir-Goudet estimator measures how much a given pair departs from a *random pair* comparison and, as such, can take on negative values. The latter simply means that is pair is less similar than an average comparison in the sample. This estimator does not attempt to estimate base-population allele frequencies, and simply takes the base population as the sample, which could contain a significant fraction of inbred individuals and/or relatives. Again, for a small sample one could compute $\overline{S}$ by first removing the pair of interest.

Single-point Correlation Estimators

Closely related to IBD probability-based estimators (Equations 8.9–8.12) are estimators based on the average correlation among alleles. As mentioned, f (and by extension Θ) can be regarded as either an IBD probability or a correlation among alleles (Cotterman 1940; Malécot 1948). This equivalence of Θ with a correlation immediately suggests why some estimates can be negative, with $\Theta < 0$ implying that the two individuals are more *dissimilar* than expected by chance. This provides a complementary viewpoint to negative estimates based on IBD probabilities, which are interpreted as the IBD probability for a specific pair being less than the expected IBD probability for a random pair from the base population.

To proceed, code the two alleles at a given SNP locus (ℓ) as 0 or 1, and let the random variable b_ℓ denote the value of a randomly drawn allele from this SNP, where $E[b_\ell] = p_\ell$ is the frequency of the reference allele (1), and $\sigma^2(b_\ell) = p_\ell(1-p_\ell)$. Analogous to Equation 8.10a, the probability that a randomly drawn allele from x and y are both 1 is

$$\Pr(b_{x,\ell} = b_{y,\ell} = 1) = \Theta_{xy} p_\ell + (1-\Theta_{xy}) p_\ell^2 \qquad (8.13a)$$

which rearranges to

$$\Theta_{xy,\ell} = \frac{\Pr(b_{x,\ell} = b_{y,\ell} = 1) - p_\ell^2}{p_\ell(1-p_\ell)} = \frac{E[b_{x,\ell}\, b_{y,\ell}] - (E[b_\ell])^2}{p_\ell(1-p_\ell)}$$
$$= \frac{\sigma(b_{x,\ell}, b_{y,\ell})}{\sigma(b_{x,\ell})\sigma(b_{y,\ell})} = \rho(b_{x,\ell}, b_{y,\ell}) \tag{8.13b}$$

namely, the correlation, ρ, between a random allele in x and a random allele in y. Toro et al. (2011) recode Equation 8.13b by using the fraction of reference alleles at locus ℓ in individual x,

$$g_{x,\ell} = \begin{cases} 0 & \text{for } 00 \\ 1/2 & \text{for } 10 \\ 1 & \text{for } 11 \end{cases} \tag{8.13c}$$

to obtain the single-locus estimate of Leutenegger et al. (2003)

$$\widehat{\Theta}_{xy,\ell} = \frac{(g_{x,\ell} - p_\ell)(g_{y,\ell} - p_\ell)}{p_\ell(1-p_\ell)} \tag{8.13d}$$

where p_ℓ is the allele frequency in the base population, which is generally replaced by the average allele frequency, $\widehat{p}_\ell$, in the sample. This can introduce significant bias, especially for rare alleles (Yang et al. 2010a).

Similarly, we can consider the correlation in reference allele *copy number* between x and y. Let $T_{x,\ell}$ denote the number of copies of allele 1 (from SNP ℓ) that x carries, where $T_{x,\ell} = 0, 1,$ or 2, corresponding, respectively to SNP genotypes of 00, 10, and 11 (note that $T_{x,\ell} = 2g_{x,\ell}$). Hence, $E[T_\ell] = 2p_\ell$, yielding a contribution to the covariance in T_ℓ (the **molecular covariance**; Toro et al. 2011) between two relatives of

$$(T_{x,\ell} - 2p_\ell)\,(T_{y,\ell} - 2p_\ell)$$

The variance in T_ℓ is $\sigma^2(T_\ell) = E[T_\ell^2] - (E[T_\ell])^2$, which under, Hardy-Weinberg, becomes

$$\left[0^2 \cdot (1-p_\ell)^2 + 1^2 \cdot 2 \cdot p_\ell(1-p_\ell) + 2^2 \cdot p_\ell^2\right] - (2p_\ell)^2 = 2p_\ell(1-p_\ell)$$

yielding a correlation in T between x and y at SNP ℓ of

$$\rho(T_{x,\ell}, T_{y,\ell}) = \frac{\sigma(T_{x,\ell}, T_{y,\ell})}{\sigma(T_{x,\ell}) \cdot \sigma(T_{y,\ell})} = \frac{(T_{x,\ell} - 2p_\ell)\,(T_{y,\ell} - 2p_\ell)}{2p_\ell(1-p_\ell)} \tag{8.14a}$$

To relate $\rho(T_{x,\ell}, T_{y,\ell})$ with Θ_{xy}, we write $T_{x,\ell} = b_{x,\ell} + b'_{x,\ell}$, where b and b' represent the two SNP alleles in x, to yield that

$$\sigma(T_{x,\ell}, T_{y,\ell}) = \sigma(b_{x,\ell} + b'_{x,\ell}, b_{y,\ell} + b'_{y,\ell}) = 4\sigma(b_{x,\ell}, b_{y,\ell})$$

Substituting this result into Equation 8.14a and recalling Equation 8.13b yields

$$\rho(T_{x,\ell}, T_{y,\ell}) = \frac{4\sigma(b_{x,\ell}, b_{y,\ell})}{2p_\ell(1-p_\ell)} = 2\left[\frac{\sigma(b_{x,\ell}, b_{y,\ell})}{p_\ell(1-p_\ell)}\right] = 2\rho(b_{x,\ell}, b_{y,\ell}) = 2\Theta_{xy} \tag{8.14b}$$

Again, the issue is how to translate the single locus correlation estimators (Equation 8.14a) into a composite single estimate. Averaging over loci yields

$$\widehat{\Theta}_{xy,V2} = \frac{1}{2L}\sum_{\ell=1}^{L} \rho(T_{x,\ell}, T_{y,\ell}) \tag{8.15a}$$
$$= \frac{1}{2L}\sum_{\ell=1}^{L} \frac{(T_{x,\ell} - 2p_\ell)\,(T_{y,\ell} - 2p_\ell)}{2p_\ell(1-p_\ell)} \tag{8.15b}$$

Widely used (e.g., Hayes et al. 2009; Yang et al. 2010a, 2011a), this estimator appears to be proposed first by VanRaden (2007, 2008) and is often called **VanRaden's second method**. The concern is that SNPs with rare alleles are much more heavily weighted, with weight $\sim 1/q_\ell$ where q is the minor allele frequency. At first blush, this seems reasonable, as the sharing of rare alleles does indeed provide stronger evidence for IBD than the sharing of common alleles. However, the base-population frequencies of such rare alleles are difficult to estimate with precision, and hence these large weights have very large sampling errors. Yang et al. (2011a) proposed a regression-based correction to adjust Equation 8.15b for using estimated frequencies in place of the true frequencies.

An alternative, and potentially more robust, estimator was also offered by VanRaden (2008), namely **VanRaden's first method**,

$$\widehat{\Theta}_{xy,V1} = \frac{\sum_{\ell=1}^{L} (T_{x,\ell} - 2p_\ell)(T_{y,\ell} - 2p_\ell)}{4\sum_{\ell=1}^{L} p_\ell(1-p_\ell)} \tag{8.16a}$$

This estimator weights all loci equally, and can be thought of as the sum of the single-locus covariances in allelic copy number divided by the sum of the variances in allelic copy number. Note that all of these correlation-based estimators (Equations 8.13d, 8.15b, and 8.16a) require unbiased estimates of the base population allele frequencies and further assume that the base population is an unstructured, randomly mating population. When either of these assumptions fail, these yield potentially biased estimates. Both of the VanRaden estimators (Equations 8.15b and 8.16a) are special cases of the estimator

$$\widehat{\Theta}_{xy,s} = \frac{\sum_{\ell=1}^{L} (T_{x,\ell} - 2p_\ell)(T_{y,\ell} - 2p_\ell)[2p_\ell(1-p_\ell)]^s}{2\sum_{\ell=1}^{L}[2p_\ell(1-p_\ell)]^{(1+s)}} \tag{8.16b}$$

where $s = 0$ and $s = -1$ correspond, respectively, to VanRaden's first and second methods. The choice of s places different amounts of emphasis on SNPs with rare alleles, with $s = 0$ weighting them equally, and $s = -1$ assigning them more weight.

Example 8.5 Wang et al. (2017) showed that many of the above method of moments estimators of Θ (e.g., Equations 8.11c, 8.15b, 8.16b) can be written as special cases of a more general single-locus estimator

$$Z_\ell(a_\ell) = \frac{T_{x,\ell} \cdot T_{y,\ell} - a_\ell(T_{x,\ell} + T_{y,\ell}) + 2a_\ell p_\ell - 4p_\ell^2}{p_\ell(1-p_\ell)} \tag{8.17a}$$

Specifically, they showed that $E[Z_\ell(a_\ell)] = \Theta$ for all values of a_ℓ (when the ture allele frequencies are used). These locus-specific values are then weighted to form the composite estimate

$$\widehat{\Theta}_{xy}(\mathbf{a}, \mathbf{w}) = \sum_{\ell=1}^{L} w_\ell \cdot Z_\ell(a_\ell), \quad \text{where} \quad \sum_{i=1}^{L} w_i = 1 \tag{8.17b}$$

with $\mathbf{a} = (a_1, \cdots, a_L)^T$ the vector of locus-specific a values for Equation 8.17a, and $\mathbf{w} = (w_1, \cdots, w_L)^T$ the vector of weights for the Z_ℓ. From Equation 3.11b,

$$\sigma^2\left[\widehat{\Theta}_{xy}(\mathbf{a}, \mathbf{w})\right] = \sum_{i=1}^{L} w_i^2 \sigma^2\left[Z_i(a)\right] + \sum_{j \neq i}^{L}\sum_{i=1}^{L} w_i w_j \sigma\left[Z_i(a), Z_j(a)\right] \tag{8.17c}$$

In the absence of LD, the double sum is zero, as genotypes are uncorrelated over loci. Wang et al. found that the optimal weights that minimizes Equation 8.17c are functions of the true relatedness (e.g., Equation 8.10e). They suggested an improved estimator for Θ by using a

two-step process: First, one uses a standard method to provide initial estimates for Θ. These are then submitted into their expressions to obtain the optimal weights in Equation 8.17b to generate an updated estimate.

Usually when estimating Θ, the goal is to obtain all pairwise combinations among a collection of individuals, with the results compactly displayed in matrix form (Chapter 9), namely the **numerator relationship matrix**, **A**. This matrix arises when using BLUP, in genomic prediction, and in correcting for cryptic relationships in a GWAS (Chapters 10, 20, 31, and 32). The ijth element of **A** is $2\Theta_{ij}$ (the weight on the additive genetic variance; Equation 7.11a), and can be estimated either by using expected values from a pedigree (Chapter 7), or the realized values from marker estimates. In the latter setting, this matrix is usually denoted by **G** and is called the **genomic relationship matrix** (VanRaden 2008). Example 9.14 shows how both VanRaden estimators for all pairwise comparisons in a sample can be compactly written in matrix form. Chapter 31 examines methods for combining both pedigree and marker information, generating a **pedigree-genomic relationship matrix**, **H** (Legarra et al. 2009; Christensen and Lund 2010).

Which Single-point Θ Estimator Should be Used?

Most of the early studies examining the effectiveness of various marker-bases estimators of Θ were done in the era when just a few dozen SSRs, or up to a few hundred SNPs, were scored. Generally, these studies found that the correlation between values predicted from known pedigrees and values predicted from markers was modest. Part of this divergence is expected, because we are contrasting an expected value (pedigree) with a specific realization (markers). The latter, provided that we can estimate it with precision, is the more accurate measure of relatedness. A second source of error, that does not diminish as one adds more markers, is potentially biased corrections for AIS. Most methods use a Hardy-Weinberg correction (e.g., Equation 8.10b), which requires accurate estimates of the base population allele frequencies, and more subtly, that the base population itself is randomly mating. Typically, allele frequencies are estimated from a reference sample, which leads to biased estimates if the sample is not representative of the base or if the sample contains any relatives or inbreds (Wang 2014). These complications can be dealt with by using the reference sample as the base, and then adjusting for the value of a randomly draw pair from the reference (e.g., the Weir-Goudet estimator; Equation 8.12).

Despite these concerns, more recent studies using significantly more markers found much better performance. For example, good performance has been found using 10^3 SNPs in captively bred populations of zebra finches (Santure et al. 2010), pigs (Lopes et al. 2013), and cattle (Rolf et al. 2010), and in natural populations of soay sheep (Bérénos et al. 2014). Goudet et al. (2018) suggested that 10^4 SNPs will very efficiently estimate realized Θ values in most settings, especially when using the reference sample as the base and an AIS correction based on sample properties (Equation 8.12b rather than Hardy-Weinberg). All of this leads to the current consensus that estimating relatedness using on the order of 10^4 SNPs provides a more accurate picture of shared relatedness than a pedigree constructed without error (Wang 2016). Early contrary views (e.g., Van Horn et al. 2007) were influenced by the poor performance of estimators using only a modest number of markers.

In terms of the relative performance of specific estimators, if the population sample consists mainly of outbred and unrelated individuals, most methods perform equally well, with some methods slightly better for close relatives and others slightly better for more distant relatives. When the sample size, and/or the actual base population is small, such that both inbreds and relatives are common, then the Weir-Goudet estimator performs better, as it does not assume Hardy Weinberg and does not rely on estimates of base-population allele frequencies. Their AIS corrector, $\overline{S}$, also takes into account correlations among marker alleles at different loci, so this class of estimators is at least somewhat robust to SNP oversampling

(high correlations among SNPs). Methods that are sensitive to rare alleles, namely when single-locus results are weighted differentially (e.g., Equation 8.10f; 8.13d, 8.15b), are very sensitive to estimates of these frequencies, and can be problematic.

LD-informed Estimators

Estimators leveraging linkage disequilibrium information—correlations between alleles at different markers—fall into two distinct categories. The first approach is a straight-forward extension of combining the single-locus estimators, with the weighting scheme for the site-specific estimates being modified to adjust for correlations among marker alleles at different loci (Day-Williams et al. 2011; Speed et al. 2012; Wang et al. 2017). Equation 8.17c shows their basic variance structure, with the double sum of covariances accounting for the effects of linkage and LD. These estimates return improved (lower sampling error), and usually more robust, estimates of Θ.

The second approach uses very dense SNP data to locate shared stretches of IBD—**segmental IBD**—in distant relatives (Purcell et al. 2007; Browning 2008; Kong et al. 2008; Albrechtsen et al. 2009; S. Browning and B. Browning 2010, 2012; Gusev et al. 2009; Manichaikul et al. 2010; B. Browning and S. Browning 2011, 2013; Han and Abney 2011; Huff et al. 2011; Moltke et al. 2011; Lawson and Falush 2012; Durand et al. 2014; Li et al. 2014; Glazner and Thompson 2015; Staples et al. 2016; Ramstetter et al. 2017). In Europeans, IBD segments as small as 2 cM (roughly 2 Mb) can routinely be detected using standard SNP arrays (Browning and Browning 2012), which represent individuals with a common ancestor of around 25 generations in the past. Recall from Example 8.4 that most pairs with such distant ancestors likely share *no* IBD segments (Equation 8.8a gives this probability as 99.4%), but *when they do*, the segments are often long enough to be detectable. Applying Equation 8.8c (with the values of M and L used in Example 8.4) gives an expected size of 1.7 Mb for 25 generations, but there is considerable variation about this value, with some segments being substantially longer. The signal used to declare a region as being IBD is an anomalously long run of matching (AIS) SNP alleles across the two individuals. This requires a sufficient number of polymorphic SNPs, with the minimal segment length that can be detected being a function of the amount of polymorphism per unit of recombination. As such, the minimal size can vary dramatically over both populations and species. It can also vary dramatically *within* a genome or chromosome, as regions of low recombination (such as around centromeres) tend to have greatly reduced levels of polymorphism (WL Chapter 8).

The computational approach for detecting IBD segments is based on **hidden Markov models** (**HMMs**). Full details can be found in most of the above references, but the basic idea is as follows. One starts at a given position on a chromosome with some estimate of the site IBD status based on marker information. The IBD state at an adjacent site is then predicted using a transmission probability (based on map distance and marker information) of remaining in the current state (the adjacent site has the same IBD status as the starting site) or moving to a new state (the new site has a different IBD status). One then moves to the next adjacent site and repeats the process. The resulting output is a probability map for runs of IBD across a chromosomal region. Underpinning this sophisticated mathematics is a rather simple idea: as was the case for rare alleles, ***individuals that share rare*** (highly improbable) ***haplotypes are likely related***.

It is important to recall one of the lessons from Example 8.4, namely, that distant relatives only share short segments of IBD (if at all) that are randomly distributed over their genomes, resulting in a chimeric organization of shared IBD regions over a chromosome. Suppose that individuals x, y, and z all shared the same relative in the distant past. By chance, pairs from this group will share different IBD segments from this ancestor (if they share any at all). For example, x and y may share a segment on (say) chromosome one, but x and z may not, but rather could share a segment on chromosome three, while y and z share a segment on chromosome five. Similarly, x and w may share a different common ancestor, so that while part of chromosome one in x and y has an IBD segment from the x-y common ancestor, a nearby region in x and w may share a different IBD region from the x-w common

ancestor. Hence, when the sample contains a set of individuals with only distant ancestors, ***the genome of a focal individual is a patchwork of IBD segments***, with most of the genome not sharing any IBD blocks with others in the sample, while the rest is a collection of small segments that are IBD with different sets of sampled individuals.

Quantification of the amount of shared IBD segments between a pair of individuals provides a direct measure of the total fraction of the genome shared by the pair (total length of shared IBD segments divided by diploid genome size), and hence a value of 2Θ. While physical distance (e.g., base pairs) are often used, the more appropriate metric is the amount of recombination (e.g., cMs), as the size of IBD regions is a function of their recombination rate. This is the continuous *relatedness* measure for this pair, which is often attempted to be mapped into a discrete *relationship* category, such as 5th cousins. Several of the programs that search for IBD segments return such genealogical estimates, usually in the form of the total number, m, of meioses that separate the two focal individuals. For example, Huff et al. (2011) developed an ML estimator for m, while Purcell et al. (2007) considered the smallest value of m that is consistent with the number and length of IBD segments. Applying Equation 7.6, we can translate an estimate of m into an estimate of Θ, with $\widehat{\Theta} = (1/2)^{\widehat{m}}$. However, as noted earlier, as the common ancestor becomes more distant, both the skew and CV of the amount of IBD sharing increases, creating every-larger uncertainties about the true nature of the relationship (Hill and Weir 2011). Indeed, Speed and Balding (2015) have argued that the notion of a specific relationship has little merit in the era of whole-genome sequencing, as one can directly measure the fraction of two genomes that are shared AIS, a point we also endorse.

MARKER-BASED ESTIMATES OF THE COEFFICIENTS OF INBREEDING (f) AND FRATERNITY (Δ_{xy})

The logic used above for developing estimates for Θ extends to other relatedness metrics. We consider two here: the amount f_x of inbreeding in individual x, and the coefficient of fraternity Δ_{xy} between x and y. As above, while pedigrees can estimate expected values for these quantities (Chapter 7), with sufficiently dense marker data we can estimate the actual realizations, which can be deviate substantially from their pedigree-expected values. Similarly, correcting f_x and Δ_{xy} estimators for AIS alleles depends on assumptions about the base population.

Single-point Estimators of f

With probability f, the two alleles at a randomly chosen locus in an inbred individual are IBD. Hence, the impact of inbreeding on genotype frequencies is straightforward (Chapter 12): decreasing the frequency of heterozygotes relative to Hardy-Weinberg,

$$\text{Freq}(A_\ell a_\ell) = (1-f)2p_\ell(1-p_\ell) \tag{8.18a}$$

and correspondingly increasing the frequency of homozygotes,

$$\text{Freq}(A_\ell A_\ell) = fp_\ell + (1-f)p_\ell^2 = p_\ell^2 + fp_\ell(1-p_\ell) \tag{8.18b}$$

It is important to stress that an inbred *individual* has a genome consisting of a *mixture* of inbred and outbred *loci* (or, more accurately, *chromosomal segments*), with a fraction f that are inbred (IBD) and fraction $(1-f)$ that are not. Absent a rare mutation, an inbred locus is never heterozygous, while homozygotes can occur at both inbred (**autozygous**) and outbred (**allozygous**) loci.

The issue in estimating the fraction, f, of an individual's genome that is autozygous is separating autozygous from allozygous homozygotes. This is usually done by comparing the observed homozygosity with the value expected given some reference population. In exactly the same fashion, the deficiency of heterozygotes also provides an estimate of the

fraction of the genome that is autozygous. Both these approaches raise, once again, the critical issue of the appropriate reference population. In many settings, the reference population itself may be inbred, so that an estimate of f, as with Θ above, reflects the *relative* difference in f between the focal individual and a random individual from the reference, and thus can be negative.

A number of method of moments estimators have been proposed based around the comparison of observed versus expected homozygosity (or heterozygosity) values. We first consider single-point estimates (f for a specific site), from which a composite estimate for the genome (or chromosomal region) is constructed by some weighting scheme of the point estimates. The basic structure is as follows. Consider a focal individual (x), and (as above) let $T_{x,\ell}$ denote the copy number of reference alleles (which we arbitrarily take as the major—more common–allele, and denote it by 1) at locus ℓ in x (taking on values of 0, 1, or 2; corresponding, respectively to SNP genotypes of 00, 10, and 11). Note that $T_{x,\ell}(2 - T_{x,\ell})$ is only nonzero for a heterozygote. The expected heterozygote frequency when an individual is inbred to level f is given by Equation 8.18a. We can use this equation to obtain an estimate of f using multilocus data in several ways. First, one could simply average both sides of Equation 8.18a over L markers in the focal individual, yielding the expectation

$$E\left[\frac{1}{L}\sum_{\ell=1}^{L} T_{x,\ell}(2 - T_{x,\ell})\right] = (1 - f)\frac{1}{L}\sum_{\ell=1}^{L} 2p_\ell(1 - p_\ell) \tag{8.19a}$$

which rearranges to give the estimator

$$\widehat{f}_{HOM,1} = 1 - \frac{\sum_{\ell=1}^{L} T_{x,\ell}(2 - T_{x,\ell})}{\sum_{\ell=1}^{L} 2p_\ell(1 - p_\ell)} = 1 - \frac{O[Het]}{E[Het]} \tag{8.19b}$$

where $O[Het]$ and $E[Het] = \sum_\ell 2p_\ell(1 - p_\ell)$ denote, respectively, the numbers of observed and expected marker heterozygotes within an individual. We can equivalently consider Equation 8.19b as an estimate based on the observed excess in homozygotes, rearranging to yield (Purcell et al. 2007)

$$\widehat{f}_{HOM,1} = \frac{O[Hom] - E[Hom]}{L - E[Hom]} \tag{8.19c}$$

which follows from noting that $O[Hom] + O[Het] = E[Hom] + E[Het] = L$. Hence

$$E[Hom] = L - E[Het] = L - \sum_{\ell=1}^{L} 2p_\ell(1 - p_\ell) \tag{8.20a}$$

$E[Hom]$ and $E[Het]$ are, respectively, the expected number of homozygous and heterozygous marker loci in a random individual from a randomly mating base population. These are akin to the AIS corrections in many of the above estimates of Θ, which also critically depended on random mating and known allele frequencies. When inbreeding is present in the base population, these assumptions fail. The approach taken by the Weir-Goudet estimator (Equation 8.12) offers a robust solution to this problem. Instead of calculating the expected number of homozygotes (or heterozygotes) under random mating assumptions, we instead use the average value of these in the *reference sample*. For example,

$$E[Het] = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\ell=1}^{L} T_{i,\ell}(2 - T_{i,\ell})\right) \tag{8.20b}$$

with $E[Hom] = L - E[Het]$. For small samples, one could compute Equation 8.20b with the focal individual excluded. These values are substituted into Equation 8.19b and 8.19c, with

the estimated f now being the departure from the average base population value, which can easily be negative (less inbred than a random individual from the base population).

An alternative estimator for f follows by rearranging Equation 8.18a to give (Li and Horvitz 1953)

$$f = 1 - \frac{\text{freq}(A_\ell a_\ell)}{2p_\ell(1-p_\ell)} \tag{8.21a}$$

which yields

$$\widehat{f}_{HOM,2} = 1 - \frac{1}{L}\sum_{\ell=1}^{L} \frac{T_{x,\ell}(2-T_{x,\ell})}{2p_\ell(1-p_\ell)} \tag{8.21b}$$

Equation 8.21b differs from Equation 8.19b in that the latter weights all loci equally, while the former places more weight on heterozygotes arising from loci with extreme allele frequencies. As such, it is a more delicate estimate than Equation 8.19, being overly sensitive to rare alleles (with their corresponding estimation uncertainties).

The same logic leading to Equation 8.15 can be used to obtain an estimator based on the correlation among uniting gametes (Yang et al. 2010a),

$$\widehat{f}_Y = \frac{1}{L}\sum_{\ell=1}^{L} \gamma_\ell, \quad \text{with} \quad \gamma_\ell = \frac{T_{x,\ell}^2 - (1+2p_\ell)T_{x,\ell} + 2p_\ell^2}{2p_\ell(1-p_\ell)} \tag{8.22a}$$

where the locus-specific estimators simplify to

$$\gamma_\ell = \begin{cases} (1-p_\ell)/p_\ell & \text{for } 1_\ell 1_\ell \ \ (T_{x,\ell}=2) \\ -1 & \text{for } 1_\ell 0_\ell \ \ (T_{x,\ell}=1) \\ p_\ell/(1-p_\ell) & \text{for } 0_\ell 0_\ell \ \ (T_{x,\ell}=0) \end{cases} \tag{8.22b}$$

Because $p_\ell \geq 1/2$ (by construction, as we take 1_ℓ to be the major allele), these weights place more value on the rare homozygote ($0_\ell 0_\ell$) than on the common homozygote ($1_\ell 1_\ell$). Equation 8.22a computes how similar the alleles at SNPs are within an individual relative to randomly mating individual from the base population. As such, it is a biased estimate when the base population itself is inbred, and/or if we have a biased estimate of the base population allele frequencies. Finally, recall Equation 7.3b, $f_x = 2\Theta_{xx} - 1$, so that one can repurpose many of the above Θ estimators by simply considering the coancestry of an individual with itself.

Several authors have proposed more sophisticated ML based estimators for f, which have been extended to allow for null alleles and to jointly estimate the allele frequencies along with f (Vogl et al. 2002; Wang 2011a; Hall et al. 2012). These are computationally more feasible than ML estimates of Θ, as one only has to obtain n estimates, as opposed to the $n(n-1)/2$ pairwise estimates for Θ.

LD-based Estimators of f: Runs of Homozygosity (ROHs)

Previously, we used dense SNP data to detect blocks of shared IBD segments *between* two individuals. We extend this approach to detect inbreeding by considering stretches *within* an individual that are IBD between its two homologous chromosomes. The idea is to use **runs of homozygosity** (**ROHs**), where a run is defined as a continuous DNA segment that is completely homozygous (Broman and Weber 1999; Chapman and Thompson 2003; Purcell et al. 2007; McQuillan et al. 2008; Howrigan et al. 2011; Keller et al. 2011; Curik et al. 2014; Peripolli et al. 2016; Ceballos et al. 2018). This is the extension of the idea of autozygous versus allozygous from loci to chromosomal segments (Browning and Browning 2011 refer to an autozygous segment as **homozygosity by descent, HBD**). This leads to the simple estimate of the fraction of the genome that is autozygous as

$$\widehat{f}_{ROH} = \frac{\text{total length of ROHs}}{\text{genome size}} \tag{8.23}$$

The key issue is distinguishing between an allozygous ROH arising by chance in an outbred individual versus an autozygous ROH arising from HBD, with our intuition being that longer ROHs are more likely to be autozygous. This is often done using so-called **rule-based methods**, namely specifying a **size threshold** used to designate a homozygous region as a ROH. The delicate problem is *linkage disequilibrium*, as a sufficiently short region of DNA persists in the population as essentially a block, and thus should be weighted as a single locus. Hence, a short ROH could be consistent with allozygosity, having no more weight that a single homozygous SNP. McQuillan et al. (2008) suggested a minimal threshold size in excess of the 100kb LD block size commonly found in many human populations, whereas later authors suggested much higher threshold values. Obviously, the LD structure of different populations/species would result in different threshold levels, see Ceballos et al. (2018). At a minimum, the ROH threshold should be (at least) several times larger that the average LD block size for the genomic region of interest.

The distribution of HBD length is approximately exponential with parameter $M/(2g)$, where M is length in Morgans of the region of interest (typically, a chromosome) and g is the number of generations since the common ancestor (Fisher 1954). For example, in a region with an average recombination rate of 1 cM (0.01 Morgans) per Mb of DNA, a threshold size of 5 Mb corresponds to ancestors within the last 10 generations, while a threshold size of 0.5 Mb corresponds to ancestors within the last 100 generations. Hence, ***long ROHs represent very recent inbreeding***, while ***short ROHs* may *represent ancient inbreeding blocks*** that have yet to be broken up by recombination. The choice of a threshold thus represents a user-defined balance between the control of false positives (longer ROHs have lower false-positive rates of autozygosity) versus reduced power (ignoring shorter runs that, cumulatively, may be biologically important). This lack of a clearly defined threshold generates a certain amount of arbitrariness for ROH-based estimates of f. We examine this more in detail in Chapter 12.

A second issue is that using the physical genome size is not the appropriate metric for Equation 8.23. There is a general trend for genomic regions with low recombination rates to have very reduced variation (WL Chapter 8). Hence, large regions of the genome, such as around centromeres, show little variation, inflating the genomic fraction of ROHs (McQuillan et al. 2008). Auton et al. (2009) and Gazal et al. (2014) suggested that a more reasonable approach would be to score both ROHs and total genome size in terms of Morgans, namely *genetic distances* (expected number of crossovers), rather than *physical distances* (base pairs). Auton et al. suggested a threshold ROH value of 1 cM.

A more formal approach would be to use models that incorporate SNP-calling and sequencing errors, as well as mutation rate estimates, to infer ROHs, rather than a strict adherence to exact homozygosity throughout a long sequence. A number of authors have proposed both ML and HMM approaches to accommodate these concerns (Broman and Weber 1999; Leutenegger et al. 2003; Purcell et al. 2007; Wang et al. 2009; Browning and Browning 2010, 2013; Pemberton et al. 2012; MacLeod et al. 2013; Gazal et al. 2014; Narasimham et al. 2016; Vieira et al. 2016; Druet and Gautier 2017; Solé et al. 2017; Szpiech et al. 2017). Because estimates of f are mainly used in studies of inbreeding depression, we defer discussion of which estimator is most appropriate until Chapter 12.

Estimators of Δ_{xy}

The last relationship metric we consider is the coefficient of fraternity, Δ_{xy}, the probability that x and y share two alleles IBD (Chapter 7). In the literature, it is occasionally referred to as the **four-gene relationship coefficient**, with Θ being the **two-gene relationship coefficient**. This metric is of interest in that it is the weight on any dominance variance shared by relatives (Equation 7.11b) and it also arises in studies of inbreeding depression (Chapter 12). Equation 7.8 shows that a nonzero value of Δ_{xy} requires special circumstances, namely that the parents (m_x, s_x and m_y, s_y) of x and y must be related, as

$$\Delta_{xy} = \Theta_{m_x m_y}\Theta_{s_x s_y} + \Theta_{m_x s_y}\Theta_{s_x m_y} \tag{8.24}$$

As a result, nonzero values of Δ require that either (i) *both* the mothers *and* the fathers are

related, or (ii) the father of x and the mother of y, *and* the father of y the mother of x, are related. Such complex relationships in a sample are expected to be much less frequent than sets of relatives that simply share a single IBD allele at a given locus (e.g., Table 7.2). In any sample, the number of pairs of relatives with nonzero Δ_{xy} values is thus expected to be far less than those that have a nonzero Θ_{xy} values, resulting in lower power and less precision when estimating the former relative to the latter. Equation 8.24 also shows that the expected value of Δ scales as the square of Θ, Hence, not only are pairs of relatives with nonzero Δ expected to be much scarcer than relatives with nonzero Θ values, the resulting value of Δ is expected to be much smaller, and really only of significance between fairly close relatives.

Both methods of moments (Ritland 1996b; Lynch and Ritland 1999; Wang 2002) and likelihood (Milligan 2003; Ackerman et al. 2017) estimators of Δ_{xy} have been proposed. Our focus here is on method of moment estimators. The basic structure of likelihood estimators follows using logic similar to that for paternity assessment, see Appendix 4 (in particular, Example A4.7) for details.

For a biallelic marker (such as a SNP), the Lynch and Ritland (1999) and Wang (2002) single-locus estimator based on marker locus ℓ is

$$\widehat{\Delta}_{xy,\ell} = 1 - \frac{(4 - 4I_{xy,2} - 3I_{xy,1})\, h_\ell - 2(1 - I_{xy,2} - I_{xy,1})}{h_\ell^2} \tag{8.25a}$$

where $h_\ell = 2p_\ell(1-p_\ell) = 1 - p_\ell^2 - (1-p_\ell)^2$ is the probability that a random individual is not a homozygote at locus i. The two AIS indicator variables are defined by

$$I_{xy,2} = \begin{cases} 1 & \text{if } x = 00, y = 00 \\ 1 & \text{if } x = 10, y = 10 \\ 1 & \text{if } x = 11, y = 11 \\ 0 & \text{otherwise} \end{cases} \tag{8.25b}$$

namely, that x and y are AIS for the marker *genotypes*, while I_1 indicates that one individual is a homozygote and the other a heterozygote,

$$I_{xy,1} = \begin{cases} 1 & \text{if } x = 00 \text{ or } 11, y = 10 \\ 1 & \text{if } x = 10, y = 00 \text{ or } 11 \\ 0 & \text{otherwise} \end{cases} \tag{8.25c}$$

Extensions to markers with multiple alleles are given by Lynch and Ritland (1999) and Wang (2002). Again, this estimator assumes known allele frequencies in a random-mating base population, as an unbiased value of h_ℓ^2 depends on these assumptions holding.

As with any single-point estimator, the issue is how one weights individual locus results to obtain a single composite estimator over all the markers in the region of interest (chromosome or entire genome). One could use the Li and Horvitz (1953) scheme of weighting by population heterozyogsity, $w_\ell = h_\ell^{-1}$, giving the composite estimator for L independent loci as

$$\widehat{\Delta}_{xy} = \frac{1}{W} \sum_{\ell=1}^{L} w_\ell\, \widehat{\Delta}_{xy,\ell} \tag{8.25d}$$

with $W = \sum w_\ell$. Lynch and Ritland (1999) weight using the inverse of the sampling variance (which, in part, depends on the unknown parameters to be estimated), while other weighting schemes are examined by Wang (2002). Improved (smaller variance) composite estimators could be obtained using the approach of Wang et al. (2017), by first estimating the relationship metrics using a standard method, and then substituting these values into the Lynch and Ritland expressions for optimal weights. One could iterate of this approach until the weights stabilize.

An alternative estimator was suggest by Zhu et al. (2015), which is the dominance counterpart of Equation 8.15b,

$$\frac{1}{L}\sum_{i=\ell}^{L}\left(\frac{(x_{\ell j}-2p_\ell^2)\,(x_{\ell k}-2p_\ell^2)}{4p_\ell^2(1-p_\ell)^2}\right) \tag{8.26a}$$

where the value of $x_{\ell i}$ for locus ℓ in individual i is given by

$$x_{\ell i}=\begin{cases}0 & \text{genotype is 00}\\ 2p_\ell & \text{genotype is 10}\\ (4p_\ell-2) & \text{genotype is 11}\end{cases} \tag{8.26b}$$

In general, the sampling variance of $\widehat{\Delta}$ scales as $1/(Ln_a^2)$, where L is the number of markers and n_a the number of alleles per marker (Lynch and Ritland 1999). Using a smaller number of highly polymorphic markers (such as SSRs) can thus be more powerful that using a larger number of less informative markers. In contrast, the sampling variance for estimates of Θ tends to scale as $1/(Ln_a)$ (Lynch and Ritland 1999; Wang 2002), so that adding more markers has the same impact as adding more alleles per marker. The extra sensitivity of estimates of Δ_{xy} to the number of alleles is intuitive, as such estimates are based on the marker *genotypes* of x and y being AIS, and hence the more even alleles frequencies are, the greater the power. Given our comments that nontrival values of Δ are generally restricted to rather close relatives, SSRs can have reasonable power in detecting coefficients of fraternity.