Identity by descent in the mapping of genetic traits

E. A. Thompson

Department of Statistics Technical Report # 646

University of Washington, Seattle, WA, USA

September, 2017

Abstract

This report shows how the descent of genome from an ancestor to currently observed descendants results in *identity by descent* (IBD) in current individuals, and hence similarities in their DNA at genetic marker loci. Conversely, data on the marker genotypes of individuals provides inferences of shared descent of genome in current individuals, not just genome-wide, but in specific genome regions. Regions where shared genome accords with phenotypic similarities for a trait provide evidence of causal DNA at some location in the region. The report considers both data observed on defined pedigree structures, and data on population members whose pedigree relationships may be remote and are unknown. We take a model-based approach, deriving probabilities of IBD and likelihoods of mapping parameters, given observed genetic data. We first consider probabilities of gene IBD among individuals and across a chromosome, using either a known pedigree or a population-based model. We then consider probabilities of genotypic and phenotypic data on individuals, conditional on latent IBD. Thence IBD may be inferred from marker genotypes, combining information from multiple SNP markers. Finally, we show how location-specific realizations of IBD can be used to address questions of gene mapping. By focusing on IBD, we unify pedigree and population-based approaches.

This report was first written in 2013, as an invited chapter for a proposed text book on identity by descent. It was revised and updated in March 2017. I am grateful to two referees of the earlier version for their comments. Work on this report was supported in part by NIH grant R37 GM046255, and was completed in 2017 while visiting the University of St.Andrews, UK, as a visiting Carnegie Centenary Professor.

1 Introduction

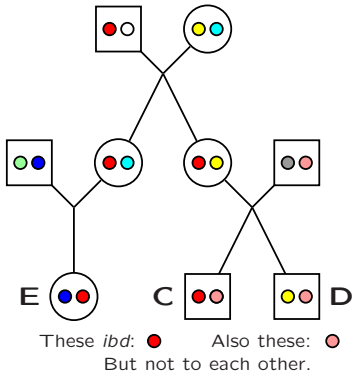
1.1 Identity by descent

Genetic similarities, whether at the population level or between close relatives, result from coancestry. Copies of DNA that descend to current individuals from a single copy of DNA in a common ancestor are, with high probability, of the same allelic type, implying greater phenotypic similarity. Such DNA is said to be *identical by descent* or IBD, and this concept is of key importance in analysis of phenotypic variation, across species of plants and animals (including humans), and across traits including disease traits, selected traits in agriculture, and normal variation.

Defining IBD is not straightforward. At every point in the genome, among any group of organisms, the coalescent ancestry will at some point converge in the *most recent common ancestor* (MRCA), and relative to this point all the organisms are IBD. Although models for this time of the MRCA of a pair of haploid genomes across genetic loci can be used for inference (Li and Durbin, 2011), for genetic mapping it will be important to have models for the changing patterns of IBD among individuals across the genome. We therefore define IBD relative to an ancestral population or time point of interest. In studies of data on defined pedigrees, IBD has often been measured relative to the pedigree founders, but these founders are members of a population and may be related or inbred relative to an earlier point in the population's history.

The choice of the time-depth of interest for analyses in IBD will depend on the scientific question. New variants arising in the distant past descend to current individuals, and remain in linkage disequilibrium (LD) with the genetic background on which they arose. This ancient coancestry (IBD) gives rise to the patterns of LD we see in populations today, and provides information on population structure and demographic history. At the other end of the scale, IBD among close relatives may be important in analysis of family data, but generally close pedigree relationships are known. In this report we focus on IBD relative to ancestors at a time depth of 10 to 40 generations. For human populations, this is beyond the depth for which pedigree information is available, or, even if available, provides useful information on coancestry of genome. On the other hand, the last 1000 years of human history encompass a large part of the huge expansion on the human species, and has established the current patterns of genetic variation within and among local populations.

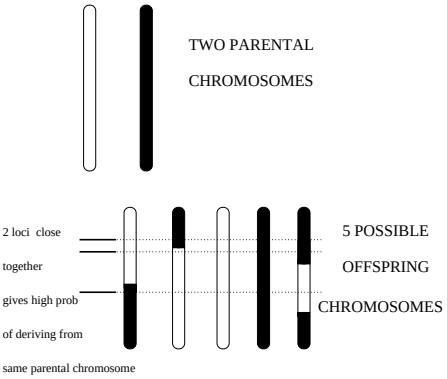
DNA is copied from parents to offspring in accordance with the process of meiosis. Throughout this report we restrict attention to nuclear autosomal chromosomes; that is, we consider neither the sex (X and Y) chromosomes, nor mitochrondrial DNA. At any given locus, the process is as specified by Mendel's first law (Mendel, 1866). A randomly chosen one of the two homologous copies of the parental DNA is copied to an offspring. A fundamental feature of the process is that this random choice is made independently in distinct meioses. Figure 1 shows a schematic view of DNA descent at a locus in a small family. At this locus, brothers C and D share their paternal DNA IBD, and cousins E



In this figure, males are designated by squares, females by circles. Individuals C and D are brothers. Individual E is their cousin, since the mothers of these three individuals are sisters. At the shown locus, copies of a single DNA (denoted by the red bullet) descends from the grandfather to cousins E and C. Additionally another single copy of DNA (the pink bullet) descends to sibs C and D from their father.

Figure 1: Descent of DNA to Sibs and Cousins

and C share their maternal DNA IBD from their grandfather.



The figure shows a pair of parental homologous chromosomes together with five examples of potential gamete chromosomes, one of which may be transmitted to an offspring.

Figure 2: Descent of a chromosome from parent to gamete

The process of meiosis also determines the copying of parental DNA across a chromosome. During the formation of gametes, in the first meiotic division, homologous chromosomes can exchange DNA leading to alternating segments of DNA from each of the two parental chromosomes (Figure 2). The points at which the offspring chromosome switches between the parent's maternal and paternal chromosomes are *crossovers*. Typically, for short chromosomes, about 50% will have no crossover, so that an intact parental chromosome is transmitted, and about 50% will have one crossover. In larger chromosomes, there may be multiple crossovers. On average, the distance between crossover points is of

order 10^8 base pairs (bp). From parents to offspring, DNA is inherited in long segments, but over multiple generations, repeated meioses break the IBD DNA in current individuals into smaller and smaller segments.

1.2 From descent to gene mapping

The fundamental framework of Genetic Epidemiology was formalized by Elston and Stewart (1971), who defined the three components of a genetic model: population, transmission, and penetrance. DNA variation in a population results from evolutionary processes and demographic history, arising via mutation, and modified by selection and random genetic drift. The parameters of the population model are allele, genotype and haplotype frequencies, which are often assumed known in genetic epidemiological studies. The model for transmission of DNA from parents to offspring is provided by the process of meiosis, as summarized in Mendel's First Law and the recombination probabilities across the genome (Section 1.1). The parameters of the population process normally relate only to the genetic map that provides recombination probabilities between any two loci, although it could also include models of segregation distortion or genetic interference.

Finally, the penetrance model defines the probability relationship between an individual's genotype at the relevant locus or loci and the observable data. For genetic markers this relationship is straightforward. Normally unphased genotypes at each marker locus are directly observed, although a model for typing error may be included. For genetic epidemiological traits of interest, the penetrance model is often the least certain and most complex component of the model, and successful inference will often require careful analysis of a range of possible models.

The goal of genetic mapping is to determine the genome locations of DNA that affects a phenotypic trait of interest. This mapping relies on co-inheritance of DNA at marker loci of known location and of DNA inferred to affect the trait. At the population level, coinheritance of DNA leads to linkage disequilibrium (LD), which is the basis of association mapping. In a defined pedigree (Figure 1), the dependence in inheritance between a trait and a genetic marker provides evidence that the DNA affecting the trait is in proximity to the genetic marker locus. In between these extremes, even in the absence of known pedigree relationships, evidence that individuals of similar trait phenotype share DNA IBD in particular regions of the genome, provides evidence that these regions harbor causal loci. This is the basis of IBD-based genetic mapping.

1.3 Outline of the report

The remainder of the report is divided into three main sections. The focus is on related individuals, who may therefore share genome IBD, but the relationships among the individuals may be known or unknown. Within each section we consider both pedigree data (known relationships) and population-level data

(unknown relationships). We present many of the ideas through numerical examples, but the reader should not (unless they wish to) be concerned with the details of computations and derivations. Focus instead on the qualitative message in the numbers provided: do the results make sense? and why does a given table or result provide insight into the approach to and goals of genetic mapping?

In Section 2 we consider probabilities of the underlying IBD in related individuals. At any locus, even a small number of gametes can share IBD in many different ways. The changes in the IBD pattern across a chromosome result of recombination events in ancestral meioses adds additional complexity. Additionally, these ancestral processes have high variance. Against these complexities are the facts that, on a bp scale, IBD changes slowly across the chromosome, and that, in populations, relatively simple prior models for IBD can provide a basis for inference. Sections 2.3 and 2.4 show the importance of IBD. Given a specification of the IBD, and a penetrance model, probabilities of genotypes and phenotypes, jointly across sets of observed individuals, without further reference to the descent structure that gave rise to the IBD.

While Section 2.3 provides probabilities of marker genotype data given a pattern of IBD, in Section 3 we consider the reverse problem—the inference of IBD from genetic marker data. As dense genetic marker data become increasingly available, and traits of interest become increasingly complex, there has been a shift in the paradigm of joint analysis of trait and marker data for purposes of gene mapping. Whereas models for complex traits may involve several genetic loci, each genetic marker corresponds to a single locus and simple models apply. By first analyzing the marker data to obtain patterns of IBD across a chromosome among observed individuals, direct joint consideration of marker and trait data may be avoided. Instead the patterns of IBD inferred from marker data may be used to investigate multiple trait models and hypotheses or even multiple traits observed on subsets of the same individuals (see for example, Chapman *et al.* (2015), Peter *et al.* (2016) and Saad *et al.* (2016)). Efficient methods for realizing, estimating, and storing, complex IBD summaries based on genetic marker data are key to success of this approach. This applies both in the presence of defined pedigree structures and also in populations: Section 3 considers both cases.

Finally, in Section 4 we consider approaches to genetic mapping of loci underlying phenotypes of interest, using the IBD inferred from genetic marker data. Both classical and modern approaches can be phrased in terms of IBD, and placing analyses in this framework shows there is no fundamental difference between pedigree-based and population-based approaches. Indeed, framing the problem in terms of IBD allows the combination of pedigree and population data. For close relatives, where relationships can be well-validated, the assumed pedigree is useful, not least in providing phase information on individual haplotypes. However, the location-specific IBD resulting from more remote relationships is often better inferred without reference to an assumed pedigree, and this IBD may be used in exactly the same way as pedigree-based IBD in genetic mapping algorithms. A final conclusion is thus that, with modern genetic marker data,

it is not a choice between pedigree and population data: gene mapping relies on having related individuals, but the relationships do not need to be known.

2 Probabilities of IBD

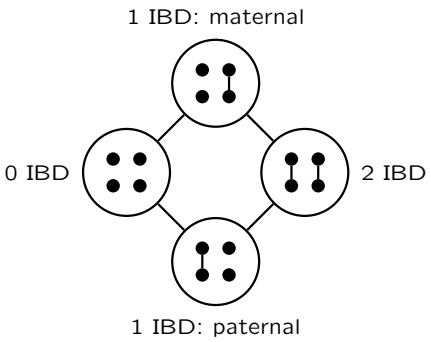
2.1 IBD in defined relatives

In a defined pedigree relationship, the probabilities of IBD are determined by the pedigree. As an example, we consider again the pair of cousins E and C in the pedigree of Figure 1. At a single locus, the probability of IBD of their maternal gametes from any one of the four grandparental genes is $(1/2)^4 = 1/16$, since that copy of the DNA must be chosen in each of four independent meioses. Of course, E and C cannot share other than their maternal genomes IBD. Thus, in total, the probability that E and C share DNA IBD at a locus is $4 \times (1/16)$ or $1/4$.

Suppose that, at a particular locus, DNA in one of the four genomes of the shared grandparent couple does descend both to E and to C, Then E and C do share genome IBD at this locus. Moving along the chromosome, we assume, for simplicity, that recombination in any meiosis occurs at random at an average rate of 1 per 10^8bp. Since a recombination event in any one of the four connecting independent meioses will break this IBD, IBD is broken at a rate 4×10^{-8} per bp, and the expected distance until IBD is broken is 25×10^6 bp (25 Mbp), which is of order 25 centiMorgans (cM). That is, an IBD segment in first cousins has expected length 25 cM. Note however that this is different from the expected length of segment surrounding a known IBD point, since the segment will extend in both directions along the chromosome, for a total expected length of 50 cM. This apparent anomaly is the phenomenon of size-biased sampling. It well known in Statistics (Cox, 1962), but has caused some confusion in the recent IBD literature. It is important to distinguish between a randomly chosen IBD segment and the segment surrounding a randomly chosen point of IBD.

Consider now the brothers C and D. At a locus they receive their maternal DNA IBD with probability $1/2$, and independently receive their paternal DNA IBD with probability $1/2$. Thus they share both homologues IBD with probability $1/4$, neither with probability $1/4$. With the remaining probability $1/2$ they share one of their two homologues IBD. Since recombination in either of two meioses breaks a maternal IBD segment, the expected length of such a segment is 50 cM, and likewise of a paternal IBD segment. The four meioses to the sibs from their parents are independent, and every recombination in any one of the four parental meioses results in a state change. Thus, along the chromosome, the IBD state remains constant for an average of 25 cM. Switches in state occur from 1 to 0 or 2 IBD or from 0 or 2 to 1 IBD (Figure 3).

We now introduce the concept of more general states of IBD at a locus, using this same example. The probabilities of the ten possible states are shown in Table 1 and may be derived as follows. There is probability $1/2$ that C and D share their paternal DNA IBD at a locus ($C_p \equiv D_p$), and probability $1/2$



The four circles show the possible IBD states, each having probability 1/4: two of these combine as the 1-IBD state with probability 1/2. Within each circle, the four bullets designate the DNA of sibs C and D. The upper[lower] pair are the two gametes of $C[D]$. Paternal gametes are on the left of each pair; maternal gametes are on the right. The lines connecting the bullets denote IBD. The lines connecting the states denote possible transitions due to recombination.

Figure 3: The states and state-changes in a pair of full sibs

	$C_p \equiv D_p$	$C_p \not\equiv D_p$	Total
$E_m \equiv C_m \equiv D_m$	1/16	1/16	1/8
$E_m \equiv C_m \not\equiv D_m$	1/16	1/16	1/8
$E_m \equiv D_m \not\equiv C_m$	1/16	1/16	1/8
$E_m \not\equiv C_m \equiv D_m$	3/16	3/16	3/8
$E_m,\ C_m\ D_m$ all $\not\equiv$	1/8	1/8	1/4
Total	1/2	1/2	1

Table 1: Probabilities of IBD states among E, C and D at any point in the genome, given the pedigree relationship of E with her sibling cousins C and D. Here, $\equiv$ denotes IBD among the specified gametes, and $\not\equiv$ denotes non-IBD.

that they do not ($C_p \not\equiv D_p$). Also, this IBD is independent of any IBD among the maternal genomes of C, D, and E. Now, for E's maternal gamete E_m to be IBD to either of the maternal gametes C_m or D_m, the same one of the four grandparental genes that descends to E must also descend to the mother of C and D: probability $4 \times (1/8) = 1/2$. The probability this same DNA is copied to both D_m and C_m, to D_m but not C_m, and to C_m but not D_m are then each $(1/2) \times (1/2) = 1/4$, giving the first three rows of Table 1. Now also, there is total probability 1/2 of IBD between the maternal gametes C_m and D_m of C and D, so that

$$\begin{aligned} \Pr(E_m \not\equiv C_m \equiv D_m) &= \Pr(C_m \equiv D_m) - \Pr(E_m \equiv C_m \equiv D_m) \\ &= 1/2 - 1/8 = 3/8, \end{aligned}$$

and the fourth row of the table follows. The final row then follows from the known column totals. We see that even in this small example, the complexity of IBD patterns increases rapidly as more gametes are considered. Here there are just five relevant gametes, and a simple pedigree relationship, but there are

already ten possible IBD combinations.

Considering changes among the ten states of Table 1 across the genome is also more complex, despite the independence of the two meioses that determine IBD between C_m and D_m and those that relate the mother of C and D to E. Additionally the rates of moving out of a given IBD state are no longer the same. C and D will share their maternal genomes for an average length of 50 cM, but IBD with E will be more rapidly broken, because of the greater number of intervening meioses. Also, it is no longer sufficient to consider only the rate of breaking an IBD chain. For example, E_m can switch directly from being IBD with C_m to being IBD with D_m.

Despite the rapidly increasing complexity of IBD states as more individuals are considered, the specification of inheritance in a defined pedigree is straightforward. Suppose all the meioses of a pedigree are indexed by m, $m = 1, ..., M$; M is twice the number of non-founders in the pedigree, since each non-founder has a maternal and a paternal meiosis giving rise to their maternal and paternal gametes. Suppose L locations of interest across a chromosome are indexed by loci j, $j = 1, ..., L$. We define, for each meiosis m and location j

$$\begin{aligned} S_{mj} &= 1 \quad \text{if the parent's paternal DNA is transmitted} \\ S_{mj} &= 0 \quad \text{if the parent's maternal DNA is transmitted} \end{aligned} \tag{1}$$

Then Mendel's first law states that meioses m are independent, and that

$$\Pr(S_{mj} = 1) \quad = \quad \Pr(S_{mj} = 0) \quad = \quad 1/2. \tag{2}$$

Secondly, under the assumption of no genetic interference (Haldane, 1919), the crossover points in the gametes transmitted to offspring (Figure 2 occur independently and at random at rate 0.01 per cM. Then the $\{S_{mj}; j = 1, ..., L\}$ have a Markov dependence over j. This can be expressed as

$$\Pr(S_{mj} = 1 \mid S_{m',j'}, (m', j') \neq (m.j)) \quad = \quad \Pr(S_{mj} = 1 \mid S_{m,(j-1)}, S_{m,(j+1)}) \tag{3}$$

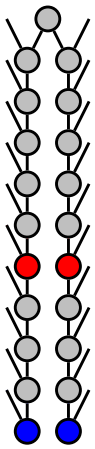
That is, given all the other $S_{m'j'}$, S_{mj}; depends only on the values $S_{m,(j-1)}$ and $S_{m,(j+1)}$ for the same meiosis m and the two neighboring loci. The vector of components S_{mj} over the values of m for any given locus j is known as the *inheritance vector* at locus j (Lander and Green, 1987).

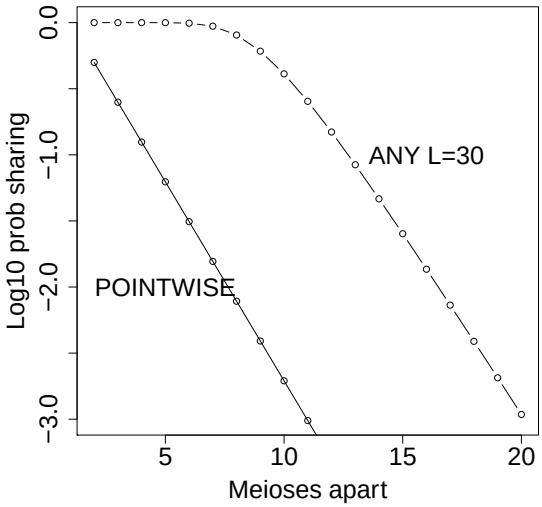
Equations (2) and (3) provide easy methods for simulation of the descent of genome in a defined pedigree. At independently inherited loci, it is simply an application of Mendel's first law, with the parents maternal or paternal DNA each being transmitted independently in every meiosis. Across the genome, the copying switches between copying from the parent's maternal and paternal DNA at a rate determined by the genetic map. Under the models of no genetic interference (Haldane, 1919), the distance to the next switch point on each chromosome can be generated as an exponential random variable with mean 100 cM. The values of S_{mj} are then determined at the specified discrete locations j. For

each j, or indeed jointly over j, the S_{mj} determine which founder genome descends to each haploid genome of each current individual, and hence the IBD state amond current individuals. Thus Monte Carlo estimates of the probabilities of IBD patterns, both at a locus and across loci, can be very efficiently obtained.

2.2 IBD in populations

In this section we consider IBD at the population level, when no pedigree relationship is specified. However, to motivate the discussion, we consider first the case of two individuals who have a single common ancestor, such that they are separated by a total of m meioses. For example, half-k^{th}-cousins have a single common ancestor $(k+1)$ generations ago and are separated by $m = 2(k+1)$ meioses.





	$m = 12$	$m = 20$
Probability of IBD at point	0.0005	2×10^{-6}
Probability of any IBD ($L = 30$)	0.148	0.001
Expected length of IBD segment	8.5 Mbp	5 Mbp

Figure 4: IBD in remote half-cousins

The probability of sharing genome IBD decreases by a factor of (1/2) with each additional meiosis. The formula for sharing any of an autosomal genome length L Morgans is more complicated (Donnelly, 1983), but also decays exponentially for larger numbers of meioses. Figure 4 shows these probabilities on a log-scale, as a function of the number of separating meioses m. At a point, the probability of IBD decays rapidly, from 0.1 at $m \approx 4$ meioses, to 0.01 at $m \approx 8$,

to 0.001 at $m \approx 11$. For a genome of length $L = 30$ Morgans, the probability of some IBD remains high for $m \leq 8$, but then starts to decrease to 0.148 at $m = 12$ and to 0.001 at $m = 20$. While 15% of pairs separated by 12 meioses will share some genome IBD from their common ancestor, this reduces to 1 in 1000 pairs for a separation of 20 meioses.

A very different picture results from considering the lengths of an IBD segment, given that it exists. IBD resulting from a chain of m meioses is broken by recombination at rate proportional to m so that length of IBD segments are of order m^{-1} Morgans. Even for $m = 20$, given there is a segment of IBD, this segment is expected to be several Mbp long.

Even for a pair of individuals, the IBD of 0, 1 or 2 gametes at a location (Figure 3) are not the only possibilities. If the parents of an individual are related, then the individual is inbred, and the two gametes within an individual may be IBD at some locations. For four gametes, there are 15 possible IBD states; the left-hand columns of Table 2 show the states in pictogram form. An alternative way to specify the IBD is by specifying the partition of the four gametes into the IBD subsets, and this specification is also given in Table 2. For example, in state 7 of Table 2 the maternal gamete A_m is IBD to both gametes B_p and B_m of B. The two subsets of this IBD partition are thus $\{A_p\}$ alone, and the three IBD gametes $\{A_m, B_p, B_m\}$. One advantage of specification as a partition is that it extends to any number of individuals or set of haploid genomes.

For model-based inference of IBD from genetic data, a prior model for IBD is required; see Section 3.4. Given a pedigree, the probabilities of Mendelian segregation and the process of meiosis provide probabilities of each IBD partition or *state*. However, in the absence of a known pedigree, an alternative approach is necessary. One natural model is that of the *Ewens Sampling Formula* (ESF: Ewens (1972)), which provides a one-parameter model for partitions of an exchangeable set of gametes. The ESF probabilities are in general written in terms of a_i, the number of subsets of size i; this specification is given in the next column of Table 2. For example, in state 7, with partition $\{\{A_p\}, \{A_m, B_p, B_m\}\}$, there is one subset of size 1 and one of size 3: $a_1 = a_3 = 1$. Since, under the ESF model, the gametes are exchangeable, all IBD partitions with the same set of values of a_i must have the same probability. For example, in Table 2, states 2, 9 and 10 have the same probability since each has $a_2 = 2$, even though in state 2 the IBD is between the two gametes within each individual and in states 9 and 10 their are two pairs of inter-individual IBD.

For our purposes the probabilities are most easily parameterized in terms of β, which is the pairwise probability of IBD between any two gametes. In terms of Ewens' classical parameter θ of genetic variation, $\beta = 1/(1 + \theta)$. For the case of four gametes the relative probability of each state is also given in Table 2. Each term is normalized by the column sum $(1+\beta)(1+2\beta)$ to give the probability. Again, in the example of state 7, there is one non-IBD factor $(1-\beta)$ and two IBD factors β to link the other three gametes. The general formula for multiple gametes is beyond the scope of this report, but the interested reader may consult Tavaré and Ewens (1997). It can also be checked that every pair

State		Partition	Ewens' $\{a_i\}$	Probability	Kinship
1		$\{A_p, A_m, B_p, B_m\}$	$a_4 = 1$	$6\beta^3$	1
2		$\{A_p, A_m\}, \{B_p, B_m\}$	$a_2 = 2$	$\beta^2(1-\beta)$	0
3		$\{A_p, A_m, B_p\}, \{B_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
4		$\{A_p, A_m, B_m\}, \{B_p\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
5		$\{A_p, A_m\}, \{B_p\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	0
6		$\{A_p, B_p, B_m\}, \{A_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
7		$\{A_p\}, \{A_m, B_p, B_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
8		$\{A_p\}, \{A_m\}, \{B_p, B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	0
9		$\{A_p, B_p\}, \{A_m, B_m\}$	$a_2 = 2$	$\beta^2(1-\beta)$	1/2
10		$\{A_p, B_m\}, \{A_m, B_p\}$	$a_2 = 2$	$\beta^2(1-\beta)$	1/2
11		$\{A_p, B_p\}, \{A_m\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
12		$\{A_p, B_m\}, \{A_m\}, \{B_p\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
13		$\{A_p\}, \{A_m, B_p\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
14		$\{A_p\}, \{A_m, B_m\}, \{B_p\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
15		$\{A_p\}, \{A_m\}, \{B_p\}, \{B_m\}$	$a_1 = 4$	$(1-\beta)^3$	0

Table 2: The 15 IBD partitions at a locus, among the four gametes of two individuals. For two individuals A and B, the paternal (p) and maternal m gametes are denoted and depicted as in Figure 3 with individual A above and B below.

of gametes has IBD probability β. For example, $A_m \equiv B_p$ in states 1, 3, 7, 10

and 13 for a total probability

$$\begin{aligned}\frac{6\beta^3 + 2\times 2\beta(1-\beta) + \beta^2(1-\beta) + \beta(1-\beta)^2}{(1+\beta)(1+2\beta)} \\ = \quad \frac{\beta(6\beta^2 + 5\beta(1-\beta) + (1-\beta)^2}{(1+\beta)(1+2\beta)} \quad = \quad \beta \end{aligned} \tag{4}$$

The classical *coefficient of kinship* between two individuals A and B is the probability that gametes segregating from each of A and B are IBD at any point in the genome. The final column of Table 2 gives this probability conditional on the IBD state among the four gametes of A and B. Note there are only four possibilities. If there is no IBD between the individuals (states 2, 5, 8, and 15), the value is 0. For one between-individual IBD link (states 11, 12,13, and 14) the value is 1/4. If all four gametes are IBD the value is 1, and the remaining 6 states each gives value 1/2. From Equation (4) and its analogues for other gamete pairs, it is seen that, under the ESF model, β is both the kinship between A and B and the inbreeding coefficient of each individual (the probability of IBD between the individual's two gametes).

The IBD state among gametes changes along a chromosome due to ancestral recombination events. For a pair of gametes, Leutenegger *et al.* (2003) proposed a simple model in which potential changes occur at rate α and at a potential change point the new (possibly unchanged) state is IBD or non-IBD with probability β and $(1-\beta)$ respectively. This gives rise to an equilibrium pairwise IBD probability β and to alternating segments of IBD and non-IBD. The lengths of these segments are exponentially distributed with expectations $1/\alpha(1-\beta)$ and $1/\alpha\beta$ respectively. This model is based on the the consideration of a single chain of ancestry (Figure 4) and does not reflect a more complex situation where there are multiple paths of ancestry of varying numbers of meioses between the two gametes. However, it is flexible enough to provide a useful prior distribution for IBD.

Modeling the changes in IBD among multiple gametes along a chromosome is a challenging problem. The full *ancestral recombination graph* is too complex a model for genomewide use. Simple approximations cannot accommodate the range of changes that can occur, or fail to mimic the types of changes that do occur. For example, an extension of the model of Leutenegger *et al.* (2003), which samples from the ESF at each potential change point, would allow immediate changes from State-1 to State-15, in Table 2, or from State-9 to State-10, whereas no single ancestral recombination event could accomplish these changes.

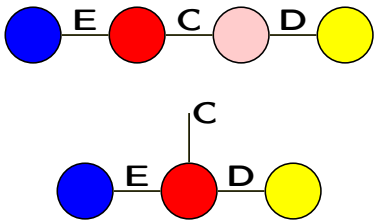
One model that has proved useful is that proposed by Brown *et al.* (2012) which applies to any number of gametes and has the ESF as its equilibrium distribution. This model allows for the move of any one gamete into, out of, or between any two IBD subsets at each potential transition point. Potential transitions occurs at rate α, which is a surrogate for recombination rate. The two parameters β and α together control the overall level of IBD, and the lengths of chromosome over which a subset of gametes will remain IBD. This model also does not accommodate all possible transitions. For example, an ancestral

recombination that is ancestral to two current gametes may move them together into another IBD subset. However, provided other changes are allowed for with some small probability, this model also provides a useful prior (Zheng *et al.*, 2014).

2.3 Probabilities of genotypic data given IBD

We now consider the relationship between latent IBD and the probabilities of marker genotypes. At a specific locus, in a specific gamete, the probability of the allelic type of the DNA is simply the population allele frequency. We will label SNP alleles as u and v, with population frequencies q and $(1-q)$. At any locus the observed genotype $G(A)$ of any individual A is thus uu, uv or vv. More generally, for a marker with k alleles ($k > 2$),i we will denote the alleles by v_i and the population frequencies by q_i ($\sum_{i=1}^{k} q_i = 1$). It is assumed that appropriate allele frequencies are known from population data bases or other sources.

The basic premise is that IBD DNA is of the same allelic type, and that non-IBD DNA copies are of independent allelic type. Mutations may cause IBD DNA to differ in allelic type, but the probability of mutation is generally much less than of typing error. We consider typing error in Section 2.4. The independence of non-IBD and the appropriate population allele frequencies are harder issues, since both depend on the frame of reference. If IBD is measured relative to a particular time-point or ancestral population, then it the allele frequencies in that population that govern the probability that a set of IBD gametes has each the same given allelic type. However, these allele frequencies are generally unknown, and instead current population estimates are used. IBD is more reliably inferred using only common genetic variants, for which population allele frequencies are more easily estimated and more stable over time.



				Q
u	u	u	v	$q^3(1-q)$
u	u	–	v	$q^2(1-q)$

$$Q = \Pr(G(E) = uu,\ G(C) = uu,\ G(D) = uv)$$

Figure 5: Example of probabilities of marker genotypes given IBD

As a simple example, we consider again the small family of Figure 1 and assume the descent of IBD that is shown in that figure. That is, at the locus of interest, cousins E and C share their maternal gametes IBD from their

grandfather, and sibs C and D share their paternal DNA IBD. This is shown graphically in the upper left graph of Figure 5. In this *IBD graph*, the observed individuals E, C and D are depicted as edges, and the notes denote the DNA. Where two or edges impinge on a node, those individuals shared that DNA IBD. Each edge joins the two DNA nodes that represent the two gene copies carried by that diploid individual.

Consider the probability that, at the marker locus of interest, E and C each has the homozygous genotype uu, while D is a heterozygote, uv. It is immediately clear that the three nodes (blue, red and pink) in E and C are of type u, while the remaining (yellow) node of D is of type v; (see right part of Figure 5). The probability that any node is type u is q, the population frequency of allele u, and likewise $(1-q)$ for v. In the total of four nodes there are 3 of type u and 1 of type v for a total probability of $q^3(1-q)$.

Now suppose it is inferred that, owing to some previously unspecified relationship between the grandfather and the father of the sibs C and D, the DNA shown as red and as pink are in fact IBD. The IBD graph is simply modified by merging the red and pink nodes (Figure 5, lower left). Individual C now carries two copies of the sane (red) DNA node, which is shared also with E and D. There are three nodes in total, two of type u and one of type v, for a total probability $Q = q^2(1-q)$.

While these examples are very simple they make important points. First, once the IBD graph and allele frequencies are given, the joint genotype probabilities are easily determined. The pedigree structure or population history that gave rise to the IBD graph is no longer relevant once the IBD graph is known. Second, at any given marker, only observed individuals are included in the IBD graph. Unlike classical pedigree computations, there is no need to sum out over possible genotypes of unobserved individuals. Third, the assignment of allelic types to the DNA nodes is trivial. If any individual is homozygous, then the only possible assignment (assuming there is one) is immediately determined by extending from that initial base. If all individuals are heterozygous there may be two alternate assignments whose probabilities must be summed. For example, a single heterozygous uv genotype not sharing DNA nodes with other individuals has genotype probability $q(1-q) + (1-q)q = 2qq(1-q)$, since either allele may be assigned to either node. Regardless of the number of possible alleles at a locus and regardless of the complexity of the IBD graph, there are always 0, 1 or 2 possible allelic assignments each connected component of the graph. Of course, on separate components we simply take the product of the probabilities on each.

2.4 Probabilities of phenotypic data given IBD

In this section, we extend the above ideas to phenotypic data. Where a phenotype allows for several underlying genotypes it is necessary to sum over the possible assignments of allelic types to DNA nodes. This computation may be accomplished sequentially across the graph, using methods that are standard in the area of graphical models (Lauritzen, 1992). Suppose for example, individu-

als E, C and D have (discrete or continuous) phenotypes Y_E, Y_C and Y_D, and that a genetic model for the trait provides the probabilities of each individual's phenotype given the unobserved allelic types of his/her DNA at the trait locus. Additionally, these penetrance probabilities may be depend on other observed covariates such as the age and gender of each individual. Them we may compute the probability of the observed phenotypes as

$$\Pr(Y_E, Y_C, Y_D) = \sum_{\bullet} \sum_{\bullet} (\Pr(Y_E|\bullet,\bullet) q(\bullet) q(\bullet) \sum_{\bullet} (\Pr(Y_C|\bullet,\bullet) q(\bullet) \sum_{\bullet} (\Pr(Y_D|\bullet,\bullet) q(\bullet)))) \quad (5)$$

where $q(\bullet)$ denotes the population allele frequency of the allelic type of DNA node $\bullet$. That is, proceeding from right to left, we may first use the information on individuals D and sum out over the possible allelic types of node $\bullet$ for each value of the type of node $\bullet$. Then we can incorporate the data on individual C, and for each value of $\bullet$ sum out over the possible values of $\bullet$. Finally we may incorporate the data on E, and sum out over the possible allelic types of the two remaining DNA nodes. By processing the summation in this way, we can break the overall sum into smaller feasible computations.

Although our example is of a small pedigree, this is for ease of presentation only. Even for more much larger and more complex IBD graphs, these computations are generally feasible, Where IBD graph components are small, probabilities can be computed even under models for which the trait is controlled by genotypes at several loci. Moreover, once the IBD graph is given, the source of that IBD information is irrelevant. The IBD graph contains all the relevant information on the impact of shared ancestry on joint phenotype probabilities.

	Non-IBD	IBD
$v_i\ v_i$	q_i^2	$(1-\varepsilon)q_i + \varepsilon q_i^2$
$v_i\ v_{i^*} (i < i^*)$	$2q_i q_{i^*}$	$\varepsilon 2 q_i q_{i^*}$

Table 3: The probabilities of an individual's genotype at a k-allele locus.. Any two distinct possible alleles at the locus are denoted v_i and v_{i*} $(1 \le i < i^* \le k)$. The population frequencies of alleles v_i and v_{i*} are q_i and q_{i^*} respectively.

A special case of phenotypic data arises with marker data where an allowance is made for typing error. That is, the true marker genotype provides probabilities for the marker phenotype; the observed marker "genotype" may not be the true one. In practice, it is important to use a model that allows for the possibility of error, so that IBD nodes are not of necessity of the same allelic type, It is not necessary to have a model that precisely reflects biological or technological genotyping processes. One simple error model for single genotypes is due to Leutenegger *et al.* (2003). In the case of IBD, with probability $(1-\varepsilon)$ the alleles are the same, and of type v_i with probability q_i, but with probability ε

the Hardy-Weinberg frequencies are used (Table 3). This allows for heterozygous genotypes even in segments where the individual's two gametes are IBD, whether due to typing error, mutation, of other causes.

For larger numbers of individuals, one error model that makes probability computations on an IBD graph straightforward is a generalization of the model of Leutenegger *et al.* (2003). With an error parameter ε, the probability of genotypes $\mathbf{g}$ is modeled as

$$\Pr(\mathbf{g}|\varepsilon, \text{IBD graph}) = (1-\varepsilon)\Pr(\mathbf{g}|\varepsilon = 0, \text{IBD graph}) + \varepsilon \Pr(\mathbf{g}|\text{ no IBD}) \quad (6)$$

That is on any connected component, with probability $(1-\varepsilon)$ there is no error, while with probability ε there is some error, and then the probability is computed as if there is no IBD among the individuals represented in that IBD-graph component.

For small numbers of individuals, or low levels of IBD, the model of Equation (6) works well, but in some cases more complex models are needed. For SNP genotypes, another model is that each allele is independently toggled to its alternative with probability ε (Zheng *et al.*, 2014). Since some loci are more error-prone, ε may be made locus-dependent. With this or a more general error model, genetic marker become in effect discrete trait phenotypes. Then probability computations require the general summation method exemplified in Equation (5).

3 Inferring IBD in pedigrees and populations

3.1 IBD given marker data on relatives

In this section we will consider the inference of IBD from genetic marker data. We consider first the case where the pedigree structure of the observed individuals is known, and assumed correct. Marker genotypes for some individuals for some subsets of loci may be missing, but we assume that, if observed, the marker genotypes are without error. As an example of the principles involved, we consider the example of a pair of full sibs. At any locus, sibs share their maternal/paternal genome IBD with probability 1/2. In the absence of genetic marker data there are prior probabilities 1/4, 1/2, and 1/4 (respectively) that they share 0, 1, or 2 gene copies IBD (Sestion 2.1).

More generally, we will denote genetic marker (usually SNP) data by $\mathbf{X}$ and a specification of the IBD to be inferred by $\mathbf{Z}$. Section 2.3 showed how probabilities $\Pr(\mathbf{X} \mid \mathbf{Z})$ of genetic marker data on relatives could be easily computed given the a pattern of IBD at the marker locus. Conversely, given a prior probability $\Pr(\mathbf{Z})$ of IBD and genetic marker data, conditional probabilities of IBD can be obtained. By Bayes theorem:

$$\Pr(\mathbf{Z} \mid \mathbf{X}) \quad \propto \quad \Pr(\mathbf{X} \mid \mathbf{Z})\ \Pr(\mathbf{Z}). \quad (7)$$

For pairs of individuals in a known pedigree relationship, there are well-established methods for computing these prior probabilities (Karigl, 1981).

Suppose the SNP genotypes of the pair of sibs at three linked loci are as shown in Table 4. As in Section 2.3 we will denote the SNP alleles as u and v, while q_j and $(1 - q_j)$ will now denote the frequency of u and of v at locus j. The reader should not struggle with details of the computation, but consider only whether the results make qualitative sense. Given the allele frequencies shown, then the probabilities that the sibs share 0, 1 or 2 DNA copies IBD are computed using equation (7) and are given in Table 4. Note the effect of the u allele frequency, q_j. The two sibs have the same genotypes at locus-1 and locus-3, but at locus-3 the u allele is the rare allele, giving much stronger weight to IBD sharing between the two sibs. Only at locus-2 do the two sibs have the same genotype, and so can share 2 copies IBD ($Z = 2$). The genotypic data raises the probability that they do so to 0.4, which is higher than above the prior probability 0.25.

			$Z=0$	$Z=1$	$Z=2$	$Z=0$	$Z=1$	Z=2
		Pr(Z)				0.25	0.5	0.25
loc-j	q_j	Data X_j	Pr($X_j \mid Z_j$)			Pr($Z_j \mid X_j$)		
loc-1	0.9	$uu,\ uv$	0.146	0.081	0.000	0.474	0.526	0.000
loc-2	0.5	$uv,\ uv$	0.250	0.250	0.500	0.200	0.400	0.400
loc-3	0.1	$uu,\ uv$	0.002	0.009	0.000	0.091	0.910	0.000

Table 4: Single-locus probabilities of IBD in a sib pair for the example data

However, if the loci are linked this computation does not take into account all the information available; there is dependence in the IBD state Z at linked loci. Suppose the recombination fraction between adjacent loci 1 and 2, and 2 and 3, are each $\rho = 0.05$. Then there is no change in the maternal [paternal] DNA sharing between adjacent loci if either both or neither of the meioses from the mother [father] to the two sibs is recombinant. The probability of this is $c = (1 - \rho)^2 + \rho^2 = 0.905$. This leads to the probabilities of transition between the IBD states $Z = 0, 1, 2$ from one locus to another at recombination distance ρ. The matrix of these transition probabilities is shown in Table 5. It can be checked that, for any value of c, this matrix has the required equilibrium single-locus probabilities 1/4, 1/2, 1/4 for $Z = 0,\ 1,\ 2$.

$Z=$	0	1	2	0	1	2
0	c^2	$2c(1-c)$	$(1-c)^2$	0.819025	0.17195	0.009025
1	$c(1-c)$	$1-2c(1-c)$	$c(1-c)$	0.085975	0.82805	0.085975
2	$(1-c)^2$	$2c(1-c)$	c^2	0.009025	0.17195	0.819025

Table 5: Transition probabilities in sib IBD states
between two loci at recombination probability $\rho = 0.05$.

We can now compute the probability of the IBD state at locus-2, given the data at all three loci. Let Z_j denote the IBD state at locus j and X_j the SNP

genotypes at locus j. Note that the data at a locus depend only on the IBD state at that locus. Thus, for example X_1, X_2, and X_3 are conditionally independent given the IBD state Z_2 at the middle locus. Then

$$\begin{aligned}
\Pr(X_1, X_2, Z_2) &= \sum_{Z_1} \Big(\Pr(X_1|Z_1)\Pr(Z_2|Z_1)\Pr(Z_1) \Big) \Pr(X_2|Z_2) \\
&= 0.0083, \quad 0.0099, \quad 0.0019 \text{ for } Z_2 = 0, 1, 2. \\
\Pr(X_3|Z_2) &= \sum_{Z_3} \Big(\Pr(X_3|Z_3)\Pr(Z_3|Z_2) \Big) \\
&= 0.0030, \quad 0.0076, \quad 0.0016 \text{ for } Z_2 = 0, 1, 2.
\end{aligned}$$

Combining these we have

$$\Pr(X_1, X_2, X_3, Z_2) \quad \propto \quad (2.490,\ 7.524,\ 0.304) \times 10^{-5} \text{ for } Z_2 = 0, 1, 2.$$

and normalizing these gives the probability of Z_2 given X_1, X_2, X_3 as approximately $(0.24, 0.73, 0.03)$ for $Z_2 = 0, 1, 2$. Note that whereas the data at locus-2 increased the relative probability that $Z_2 = 2$, incorporating the data at loci 1 and 3 greatly decreases the probability since the state of 2 IBD is impossible at each of these flanking loci. For $Z_2 = 2$, a recombination would be required both between locus-1 and locus-2 and between locus-2 and locus-3.

3.2 Monte Carlo realization of IBD in defined pedigrees

The principles that underlie the computations of Section 3.1 are that the data at each marker locus depends only on the latent IBD state at that locus, and that the transitions in latent IBD state follow a Markov process across the chromosome. This is the classic framework for a *hidden Markov model* (HMM) which enables many computations to be made. In fact, IBD is in general not Markovian, since many different patterns of inheritance may give rise to the same IBD state among observed individuals. However, in the absence of genetic interference, the indicators of maternal or paternal transmission in a meiosis are indeed Markov (See Equation (1)). This Markov nature of *inheritance vectors* (Section 2.1) has been used by many in enabling computations on pedigrees. For example, as implemented in the MERLIN software (Abecasis *et al.*, 2002), probabilities of IBD pairwise among the members of a small pedigree, may be computed at each marker locus, conditional on the joint marker data on all pedigree members and on marker genotypes and all loci. More generally, as seen in Section 2.1, the inheritance vector at a locus determines the IBD state at that locus, among all members of the pedigree.

While location-specific probabilities of IBD are useful, there are good reasons to prefer Monte Carlo realizations of latent inheritance vectors or IBD. First, it is not feasible to consider probabilities of IBD jointly across multiple loci, but a set of realizations directly display the segmental nature of inheritance and can identify locations of recombination events that change the IBD among observed individuals. Second, the variation in a set of realizations provides a

measure of uncertainty in the IBD information which cannot be captured in a single probability. In using IBD inferred from marker data in subsequent genetic analyses, it is important to have some measure of this uncertainty.

The same HMM methods that allow computation of IBD probabilities also allow Monte Carlo realization of IBD conditional on genetic marker data (Thompson, 2000). On small pedigrees, where exact computation is feasible, independent realizations may be generated. On larger pedigrees, where the space of inheritance vectors at each locus is too large for exact computation, Markov chain Monte Carlo (MCMC) methods must instead be used (Sobel and Lange, 1996; Thompson, 2000). However, the same principles apply. The Markov dependence of inheritance vectors, and the dependence of marker genotypes only on the inheritance at that marker locus, enable realizations of inheritance jointly across loci and among individuals to be made efficiently. The only requirement is that, at each marker location j, $\Pr(\mathbf{X}_j \mid \mathbf{Z}_j)$, can be easily computed (see section 2.3). However, this requirement does generally impose the restriction that marker genotypes are assumed to be observed without error.

If multiple realizations of inheritance across a chromosome are to be realized and used in subsequent genetic analyses, it is necessary to store them compactly. Note than in any meiosis, crossovers (switches between transmission of maternal and paternal DNA) occur on average only every 10^8 bp. Thus, rather than storing inheritance vectors at each location, it is more efficient to store only the initial value and the bp locations of successive switches. The inheritance vector at any location may then be efficiently determined, and consequently the IBD graph among individuals observed for a trait of interest. Only the IBD graph is relevant to subsequent analyses.

3.3 Inference of realized kinship or relatedness

A pedigree provides a very strong prior on probabilities of IBD at a locus (Section 2.1), but as genetic marker data become more and more informative, this prior is increasingly unnecessary. Moreover, for more remote relatives, IBD is highly variable. In the example of Figure 4, only 1 in 1000 pairs of individuals separated by 20 meioses will share any autosomal IBD, but if they do they will share (on average) 5 Mbp. Other examples are considered by Donnelly (1983), while Hill and Weir (2011) give a more extensive review of the variation in realized proportions of genome shared given different patterns and degrees of pedigree relatedness.

Therefore, with the current availability of dense SNP marker data, there has been an explosion of interest in the recent literature in the estimation of realized kinship from genotypic data. More often this is phrased in terms of *realized relatedness*, or of the proportion of genome shared IBD by pairs of individuals, but this is simply twice the realized kinship, which is, in turn a function of the realized 4-gamete IBD states across the genome (Table 2).

The most widely used measure of realized relatedness based on genotypes is the *genetic relatedness matrix* of GRM (see for example Hayes *et al.* (2009). The GRM is estimated as follows. As previously, at any SNP locus j, we have alleles

u and v with frequencies q_j and $(1-q_j)$. The genotype x_{ij} of an individual i can be specified by the number of u alleles he carries: $x_{ij} = 2, 1, 0$ for genotypes uu, uv, and vv respectively. Under a model of sampling alleles from the current population, x_{ij} has expectation $2q_j$ and variance $2q_j(1-q_j)$. For two individuals i and k, the (i,k) entry of the GRM is the empirical correlation between the allele counts x of i and k:

$$A_{ik} \quad = \quad \frac{1}{L}\sum_{j=1}^{L}\frac{(x_{ij}-2q_j)(x_{kj}-2q_j)}{2q_j(1-q_j)} \tag{8}$$

where L is the total number of loci genotyped. A more general form is

$$A_{ik} \quad = \quad \sum_{j=1}^{L} w_j \, \frac{(x_{ij}-2q_j)(x_{kj}-2q_j)}{2q_j(1-q_j)} \tag{9}$$

With $w_j = 1/L$ we obtain the previous estimator, while
$w_j = p_j(1-p_j)/\sum_{l=1}^{L} p_l(1-p_l)$ provides a form that is more robust to small allele frequencies; see for example VanRaden (2008).

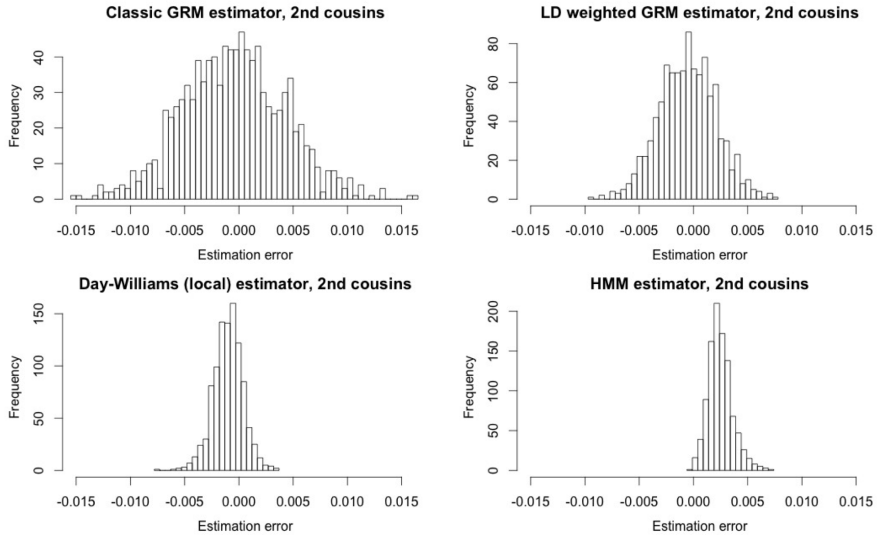


Figure 6: Histograms of estimation errors of four estimators on 1000 simulated second cousin pairs. The values are the difference between the estimated global realized kinship, and the actual (simulated) global realized kinship, computed at 169,751 SNP marker positions. The estimators are the classic GRM (8), an LD weighted version of the form (9), and estimates based on the local DW approach (DW), and on the HMM method (Section 3.4). The figure is due to Bowen Wang, based on the study by Wang *et al.* (2017)

One major deficiency of estimators where the weights depend only on allele frequencies is that they do not account for allelic associations among loci

(LD). One form of weighting to accommodate LD was developed by Speed *et al.* (2012) while S.Sverdlov developed an alternative approach that is used in Wang *et al.* (2017). The top two panels of Figure 6 show the increased precision of estimation of realized kinship (relatedness/2) by accounting for LD. The left-hand panel is for the GRM estimate of realized kinship, $A_{ik}/2$ where A_{ik} is as in Equation (8). The right-hand panel shows the results for Equation (9) with weights w_j computed according to the LD-weighting developed in Wang *et al.* (2017). Shown are simulation results for 1000 pairs of second cousins, and the histograms are of the difference between the estimated realized kinship and the actual realized kinship in each pair. Note that we are not here attempting to estimate the pedigree kinship. The histogram of differences between the pedigree value and actual realized values has a larger spread than even the upper left histogram based on Equation (8).

There is however a more serious deficiency in estimators of the form (9), and that is that they do not take the physical locations of SNPs into account. We have already seen that IBD occurs even in remote relatives as a few long segments. Additionally, SNPs are individually very uninformative; information about IBD should be combined across local SNPs to provide more accurate estimates of the probability of IBD at each point in the genome. Such estimates can then be averaged across the genome to estimate the realized proportion of genome shared IBD. One such method was proposed by Day-Williams *et al.* (2011) and is denoted DW. They use the four between-individual comparisons of allelic sharing at loci in windows across the genome, to obtain estimates of local kinship. These are then smoothed across the genome, subject to constraints that at each point the value is 0, 1/4, 1/2 or 1 (Table 2). An alternative is to estimate the IBD state for the four gametes using an HMM approach and the model of Brown *et al.* (2012) for changes in the 15 states across the genome (Section 2.2): details are given in Section 3.4 below. This method provides estimates of realized kinship at points across the genome, and these may then be combined into a genome-wide estimate.

The two lower histograms of Figure 6 show the results using the local IBD estimation methods of Day-Williams *et al.* (2011), denoted DW, and of Brown *et al.* (2012) denoted HMM. It is seen that incorporating the segmental nature of DNA into the inference process greatly improves the precision of estimation of realized kinship. However, these methods are more computationally intensive, and also show bias. The DW method tends to underestimate IBD, especially in the presence of inbreeding. The HMM method tend to overestimate IBD in the presence of LD. For this reason, an LD-weighted version of the HMM estimator provides further improvement. These and other estimators are further discussed by Wang *et al.* (2017).

3.4 IBD given marker data in populations

In the previous section the focus was on estimating the genome-wide proportions shared IBD between two individuals. However, for gene mapping, we may be interested in the joint pattern of IBD among several observed individuals, not

only pairwise measures. Second, for mapping we are interested in IBD at specific locations across the genome. Third, we may wish to consider segments of IBD, and changes in IBD across genome locations. Estimates of relatedness such as Equation (9) do not take into account the physical linkage among loci, treating them as an exchangeable collection of SNPs; any permutation of the SNPs will provide the same result. By contrast, estimates of location-specific IBD rely on the genetic marker map, and depend jointly on the SNPs in the genome region. Each SNP alone provides little evidence, but segments of IBD typically encompass many SNPs.

A defined pedigree provides prior probabilities of IBD among individuals at a locus and across a chromosome (Section 2.1). However, population genetics also provides probabilities of coancestry and IBD for individuals sampled from a population. In the context of modern highly informative SNP data, the prior distribution has relatively less weight, and genetic marker data can provide strong evidence of segments of IBD among individuals not known to be related. Formerly, where marker data were sparse both in the genome and among individuals, the highly informative pedigree prior was a necessity for successful inference. With modern data, it is often unnecessarily constraining. Further, ancestral pedigrees may be inaccurate and cannot be validated from current genetic data. Except among the current generations of sampled individuals, where the genetic data may be used to validate the pedigree, the use of a pedigree prior is often best avoided.

Model-based inference of IBD requires allele frequencies, and the relevant allele frequencies are those in the reference population relative to which IBD is measured. Sharing of a rare variant allele among individuals provides strong evidence of IBD, but on average common allelic variation provides more information. Moreover, for a rare variant, it is difficult to either quantify the strength of the evidence, or to assess its uncertainty; even the concept of a population allele frequency may be problematic. In contrast, common SNP variation is ancient and relatively stable. While each SNP alone provides little information, segments of IBD generally encompass large numbers of SNPs. Whether or not date are phased, and whether or not local haplotype frequencies are used in estimation, it is the combination of data from multiple contiguous SNPs that provides the evidence of IBD.

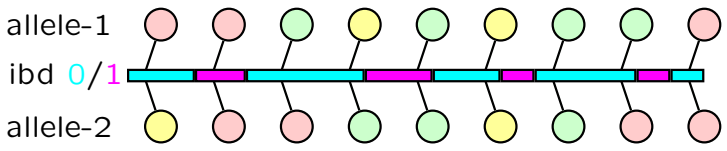


Figure 7: Model for inferring IBD between two gametes

As described in Section 2.2, a flexible 2-parameter prior model for IBD between the pair of gametes of an individual was introduced by Leutenegger *et al.* (2003). In this case there are just two possible IBD states at each locus; the two gametes are IBD ($Z = 1$) or they are not ($Z = 0$). The assumption of

a Markov process of transitions between the two states again gives an HMM structure that allows inference of IBD segments. The model for IBD is shown schematically in Figure 7. The latent IBD state consists of alternating segments of IBD ($Z = 1$) and non-IBD ($Z = 0$) between the two gametes. In an IBD segment, the allelic types at marker loci are, with high probability, of the same allelic type. In non-IBD segments they are of independent allelic types. More precisely, the data model was given in Table 3 in Section 2.4. At each marker locus, given non-IBD, we have Hardy-Weinberg genotype probabilities. In the case of IBD a small "error" probability ε allows for the possibility that IBD DNA may be, or be recorded, as of different allelic types. While the exact form of the error model is not important, it is important to allow this flexibility: generally, zero probabilities of latent states should be avoided in modelling data. Since the model has an HMM structure, standard algorithms give probabilities $\Pr(Z_j \mid \mathbf{Y})$, of the IBD state Z_j at each locus j, given the data $\mathbf{X}$ of the allele types on the gametes over all loci. Further, as in Section 3.2, we may instead obtain realizations of IBD $\{Z_j; j = 1, ..., \ell\}$ given $\mathbf{X}$, jointly over j.

There have been a number of extensions of the basic model, including to analyses of genotypes of pairs of individuals as implemented in the well-known PLINK software (Price *et al.*, 2006). A more general prior model for IBD across the genome and among IBD among any number of gametes was given in Section 2.2. This model was used by Brown *et al.* (2012) on sets of four gametes, using either haplotypic or genotypic data, and by Zheng *et al.* (2014) for multiple gametes but assuming haplotypic data. Moltke *et al.* (2011) also proposed a model for multiple gametes, and used it to study IBD in a set of 5 individuals. All these approaches have the same basic framework and objective. That is, patterns of IBD across a chromosome among sets of gametes are to be inferred from genetic marker data. The IBD process is approximated by a Markov process, and the allelic of genotypic data at each locus depends only on the underlying IBD state, giving rise to an HMM.

There is one significant approximation in these models in that linkage disequilibrium (LD) is ignored. That is, there is no direct dependence of allelic types between loci. While it is the allelic similarity of gametes across multiple loci that results in inference of IBD, haplotypic similarities are not modeled directly. Allele frequencies are incorporated into the model, and for common SNP variation these are normally adequately accurately known, but haplotype frequencies are often less well established. While ignoring LD is a model mis-specification that can result in false inferences of IBD (Figure 6), over-compensation for LD can lead to failure to detect IBD (Brown *et al.*, 2012). A model that does include LD in the inference of IBD is that due to Browning and Browning (2010), implemented in the BEAGLE package, but at the expense of a simplified IBD model. This approach works very well in large samples from large populations, where IBD levels are low and haplotype frequencies can be well-estimated.

These methods also all face another issue: as the number of gametes n increases, the number of possible IBD states at each locus increases very rapidly, being the number of partitions of n items. For the 12 gametes of 6 individuals there are more than 4 million possible states. The example considered by Moltke

et al. (2011) was for just 5 individuals and a limited gene region. Zheng *et al.* (2014) considered 860 SNP markers over a region of 10 Mbp, and succeeded in realizing joint IBD among 40 gametes, but assumed the availability of haplotypic marker data. Neither of these approaches is scalable to chromosome-wide inferences of joint IBD among multiple gametes from genotypic marker data.

4 IBD-based genetic mapping

Any genetic mapping procedure aims to detect the genomic locations of DNA variation underlying a trait of interest, by reference to a genetic map of markers that have know locations in the genome. An association test directly considers the dependence between marker genotypes ($\mathbf{X}$) and trait phenotypes ($\mathbf{Y}$). However, these allelic associations arise from the descent of DNA from common ancestors to different individuals within a population or to different populations. It is therefore useful to consider the associations between $\mathbf{X}$ and $\mathbf{Y}$ through the lens of descent $\mathbf{Z}$, and specifically patterns of IBD among individuals observed for the trait inferred at locations across the genome. Throughout this section we assume that $\mathbf{Z}$ contains all the information needed for analysis of association between $\mathbf{X}$ and $\mathbf{Y}$: that is, we assume that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$.

4.1 Mapping from IBD in pedigrees

We first consider the pedigree context, in which prior probabilities of IBD are provided by the specified pedigree relationships among individuals. At locations across a chromosome, the genetic marker data $\mathbf{X}$ provide probabilities of IBD, or realizations of the *inheritance vectors* which determine the location-specific descent of DNA through a pedigree (Section 3.2).

In genetic mapping, the use of location-specific IBD in affected individuals of known pedigree relationship has long been established as a powerful approach. This may be IBD in affected sib pairs (Suarez *et al.*, 1978) or more general relative pairs (Weeks and Lange, 1988), or even the two parental gametes of individuals affected by a rare recessive trait (Lander and Botstein, 1987). In each case, individuals sharing the trait have increased probability of sharing genome IBD at causal loci. Observation of the same genome regions showing IBD across multiple pairs of affected relatives provides the linkage signal. Significance of the observations can be readily assessed, since the known pedigree relationship provides the null distribution of any IBD-based test statistic.

IBD-based tests for linkage have been extended to larger groups of relatives and a wide variety of test statistics have been developed (McPeek, 1999). Joint sharing of genome IBD by multiple affected relatives generally provides stronger evidence, since the pedigree-based prior probability of this multiple sharing event is generally smaller. An advantage of an IBD-based approach is that it is relatively robust to allelic heterogeneity: within a pedigree the affected individuals are likely to carry the same causal mutation IBD. If there are large

pedigrees with multiple affected individuals available, even locus heterogeneity is a lesser concern, since even a single pedigree can provide sufficient evidence of linkage.

In the genetic mapping of loci underlying quantitative traits (QTL), IBD-based approaches also have a long history. Haseman and Elston (1972) developed an approach to mapping QTL by considering the (negative) correlation between the squared difference in trait values in sibs and the marker-based IBD probability. More generally, the variance-component approaches to QTL mapping in the SOLAR software (Almasy and Blangero, 1998; Blangero *et al.*, 2000) use pairwise location-specific IBD probabilities computed conditionally on the pedigree structure and on observed marker data to model covariances among relatives and map QTL.

A simple version of the model for QTL detection is as follows. The vector of trait observations $\mathbf{Y}$ over the individuals is modeled as

$$\mathbf{Y} = \mu\mathbf{1} + \sigma_a\mathbf{g} + \tau_j\mathbf{w}_j + \sigma_e\mathbf{e} \tag{10}$$

Here μ is the overall mean, which may more generally include other fixed effects and covariates, $\mathbf{g}$ is a vector of genome-wide genetic (polygenic) effects, $\mathbf{w}$ is a vector of location-specific effects, and $\mathbf{e}$ is a vector of independent individual residuals. Thus $\mathrm{Var}(\mathbf{e})$ is the identity matrix $\mathbf{I}$, and $\mathrm{Var}(\mathbf{w}_j) = 2\Phi_j$ where Φ_j is the matrix of between-individual kinships at location j (Table 2). The variance of the genome-wide effect $\mathbf{g}$ is $\mathrm{Var}(\mathbf{g}) = 2\Psi$ where Ψ is the matrix of genome-wide kinships. For any pair of individuals, the term in the matrix Φ_j may be obtained for any location j from realizations of descent conditional on marker data $\mathbf{X}$ (Section 3.2). For each pair of individuals, these values may be averaged across the genome to obtain an estimate of Ψ, although in the past a pedigree-based expectation was often used for Ψ (Section 2.1).

A log-likelihood-ratio can be used to test whether there is an effect specific to any location j in the genome. The purpose of the genome-wide term ($\sigma_a > 0$) is to absorb effects of genes other than at the test location j, in order to provide greater power and precision in detecting the effect at locus j. The general model of Equation (10) can be compared to the model in which there is no effect at location j ($\tau_j^2 = 0$):

$$\ell_j \quad = \quad \log\left(\frac{\max_{\sigma_a^2,\tau_j^2,\sigma_e^2} L(\sigma_a^2,\tau_j^2,\sigma_e^2;\Phi_j,\Psi)}{\max_{\sigma_a^2,\sigma_e^2} L(\sigma_a^2,\tau_j^2=0,\sigma_e^2;\Psi)}\right) \tag{11}$$

For pedigrees with not too many observed individuals, maximization over parameters, and hence computation of the test statistic (11), is not computationally intensive.

4.2 IBD in pedigree-based likelihoods

In model-based testing for an association between genetic marker data $\mathbf{X}$ and trait data $\mathbf{Y}$, one may consider either the probability $\Pr(\mathbf{Y}|\mathbf{X})$ or the probability

$\Pr(\mathbf{X}|\mathbf{Y})$. The latter is the basis of association studies, in which individuals are selected on the basis of their trait data, $\mathbf{Y}$ (for example, cases and controls). Genotypes $\mathbf{X}$ are then compared between these two groups. For a quantitative trait, or when a more general trait model is desired, it is more natural to consider $\mathbf{Y}$ given $\mathbf{X}$. The same applies to IBD-based genetic mapping, where the marker data $\mathbf{X}$ are used to provide information about IBD, $\mathbf{Z}$. In the analogue of population-based association tests we consider differences in inferred IBD in pairs conditional on their trait status (see Section 4.3 below). In QTL mapping we model $\mathbf{Y}$ given the inferred IBD as in Equation (10) above. In this section, we consider more generally the probability $\Pr(\mathbf{Y}|\mathbf{X})$ in the case where the pedigree structure is known.

As in classical linkage analysis Smith (1953); Morton (1955), the goal is to test for dependence between inheritance of DNA at specific genome locations and the inheritance of DNA underlying a trait The pedigree relationships among individuals are assumed known, and genetic marker data $\mathbf{X}$ are available for some individuals, for markers with known locations in the genome. Trait data $\mathbf{Y}$ are also available, and a model is assumed for the relationship between the allelic type of latent causal DNA and the trait of interest. In the following, Γ_X will denote the probability model for the marker data $\mathbf{X}$, which involves marker allele frequencies and locations which are assumed known, and Γ_Y will denote the model for the trait, which specifies frequencies of trait alleles, and the probabilities of phenotypes given latent trait genotypes. The parameter λ is a set of locations at which, with any model, there is hypothesized to be causal DNA. The full model is $\Gamma = (\Gamma_X, \Gamma_Y, \lambda)$. The full set of all locations at which causal DNA is hypothesized in any model to be considered will be denoted Λ; each λ is a subset of Λ. For models with a single trait locus, $\lambda = \{j\}$ and Λ is the set of j at which likelihoods are to be computed.

As before, we assume that $\mathbf{Y}$ and $\mathbf{X}$ are conditionally independent given $\mathbf{Z}$, and denote by $Z(\lambda)$ the IBD jointly at locations specified by λ. For single-locus trait models $\lambda = \{j\}$, we write $Z(\lambda) = Z_j$. Then

$$\Pr(\mathbf{Y} \mid \mathbf{X}; \Gamma) \quad = \quad \sum_{\mathbf{Z}} \Pr(\mathbf{Y} \mid Z(\lambda); \Gamma_Y, \lambda) \Pr(Z(\lambda) \mid \mathbf{X}; \Gamma_X) \qquad (12)$$

There are several issues inherent in the use of Equation (12). First, even though $\mathbf{Z}$ is required only at locations specified in λ and only among individuals observed for the trait, direct computation is infeasible, except in cases of small pedigrees and simple trait models. If the trait model involves more than a single trait locus, so λ is not a single point, computation of the joint probabilities of $Z(\lambda)$ given the marker data $\mathbf{X}$ across the chromosome is not possible. Next, even for a single hypothesized location $\lambda = \{j\}$ for the causal DNA, if there are more than three related individuals observed for the trait, the number of possible IBD states among them is too large for practical computation of $\Pr(Z_j|\mathbf{X})$.

However, a Monte Carlo approach is feasible. The conditional probability of $\mathbf{Y}$ given $\mathbf{X}$ may be re-written as

$$\Pr(\mathbf{Y} \mid \mathbf{X}; \Gamma) \quad = \quad \mathrm{E}(\Pr(\mathbf{Y} \mid Z(\lambda); \Gamma_Y) \mid \mathbf{X})$$

where the expectation is over the values of $Z(\lambda)$ given $\mathbf{X}$. The methods of Section 3.2 allow a large number, N, of patterns of IBD $\mathbf{Z}^k$ $(k = 1, ..., N)$ across the chromosome to be sampled jointly for all relevant locations j in any collection of models $\{\lambda : \lambda \subset \Lambda\}$ conditional on the joint marker data on individuals and across the chromosome. For any specific hypothesis λ, the required probability may be estimated as

$$\widehat{\Pr}(\mathbf{Y} \mid \mathbf{X}; \Gamma) \quad = \quad \frac{1}{N} \sum_{k=1}^{N} \Pr(\mathbf{Y} \mid Z(\lambda)^{(k)}; \Gamma_Y), \quad Z(\lambda)^{(k)} \sim \Pr(\cdot|\mathbf{X}; \Gamma_X). \quad (13)$$

Given the realizations of $Z(\lambda)$, the estimate requires only $\Pr(\mathbf{Y} \mid Z(\lambda))$ for each hypothesized λ. If the model for trait phenotypes involves only a single locus $\lambda = \{j\}$, and writing $Z_j = Z(\{j\})$, the probability $\Pr(\mathbf{Y} \mid Z_j)$ may be computed by the methods of Section 2.4. The IBD-graph approach outlined there can be extended to two-locus trait models (Su and Thompson, 2012).

For a single-locus trait model, Equation (13) is analogous to that first proposed by Lange and Sobel (1991), but is here phrased in terms of IBD rather than latent genotypes of individuals. Sampling and efficient storage of a collection of IBD realizations across the genome greatly facilitates analysis. Since Equation (12) separates the marker model Γ_X from the trait data $\mathbf{Y}$ and model Γ_Y, the analysis of the marker data may be performed once only. Computation of likelihoods directly from the stored IBD graphs allows these IBD graphs realized from marker data to be used not only for different hypothesized trait locations Lange and Sobel (1991) but also for different trait models, and even for different traits observed on subsets of the same set of individuals.

On a given pedigree component, many different realizations, across many loci, and of different inheritance vectors, may give rise to the same IBD graph. The probability $\Pr(\mathbf{Y} \mid Z_j^{(k)}; \Gamma_Y)$ should be computed only once for each equivalent IBD graph. Given a sample of IBD graphs, each across a chromosome, there are algorithms to determine when IBD graphs are genetically equivalent (Koepke and Thompson, 2013). This can greatly increase efficiency of the trait-data portion of the analysis, especially when the same collection of marker-based IBD graphs are to be used in analyses of multiple trait models or for data on multiple traits.

There are several other issues in the use of Equation (12). While the values of $\Pr(\mathbf{Y}|\mathbf{X}; \Gamma) \;=\; \Pr(\mathbf{Y}|\mathbf{X}; \Gamma_X, \Gamma_Y, \lambda)$ can be compared for different hypothesized values of λ, there is no baseline as to what should be expected for given sets of marker data $\mathbf{X}$. The classical Human Genetics approach has been to compare the value of (12) with the probability $\Pr(\mathbf{Y}; \Gamma_Y)$ under the same trait segregation model but in the absence of marker data. Note that this baseline "marker-free" null model is different from from the null model of QTL mapping (Equation (11)), which is widely used in the plant and animal literature (Lander and Botstein, 1987). In that case the null model is of a zero effect of the DNA at a specific genome location j $(\tau_j^2 = 0)$.

In the days before the existence of genome-wide genetic marker maps, a comparison of the marker-based $\Pr(\mathbf{Y}|\mathbf{X}; \Gamma)$ with $\Pr(\mathbf{Y}; \Gamma_Y)$ had a sound foun-

dation (Smith, 1953; Morton, 1955). However, this is less meaningful when values of λ or locations j are spread across the genome, and genetic markers are likewise distributed genome-wide. Further, the unconditional trait probability $\Pr(\mathbf{Y};\Gamma_Y)$ may not be computable for data observed on very large complex pedigrees, or for complex trait models. Second, even when easily computed, there remains the choice of Γ_Y. Trait models have a number of parameters, for each latent trait locus and genotype. Maximization over these parameters is often impractical, and the likelihood (12) is often sensitive to model choice. We return to these issues in Section 4.4 below.

4.3 Mapping from IBD in populations

Just as when the pedigree relationships among individuals are known (Section 4.1), population-based IBD mapping relies on excess IBD among individuals of similar phenotype, relative to some null model or comparison group. As an example, we consider IBD-based mapping in a case-control study (Browning and Thompson, 2012). To avoid the issues of inferring IBD from marker data, will assume that the local IBD between pairs of individuals is known with certainty.

Recall that in a simple association test the frequency of a SNP allele in N_1 cases is compared with that in N_2 controls:

$$\left(\frac{1}{2N_1} \sum_{\text{cases}} X_i \; - \; \frac{1}{2N_2} \sum_{\text{cont.}} X_i \right) \tag{14}$$

where $X_i = 0, 1, 2$ is number of alleles of specified type in i. By analogy, in an IBD-based test, we compare the frequency of IBD between M_1 case-case pairs and M_2 other pairs (case-(non-case) or (non-case)-(non-case)):

$$\left(\frac{1}{M_1} \sum_{\text{case-case}} Z_i \; - \; \frac{1}{M_2} \sum_{\text{other}} Z_i \right) \tag{15}$$

where $Z_i = 1$ or 0 as the pair does/does not share genome by descent at test location.

Just as in an association test, we must allow for population heterogeneity or structure. In an association test, there may be similarities among cases and/or among controls that are unrelated to the trait. Likewise in an IBD-based test, there may be different degrees of relatedness among cases from among controls, due to the methods of sampling or ascertainment. The average IBD scores within each group in equation (15) may be adjusted for the genome-wide average within in each group.

To assess significance a null distribution is required. Whereas in a known pedigree, Mendelian segregation provides an appropriate null distribution, in a population there is no such framework. However, just as in an association test

selec -tion	# var.	var.freq.	total freq. of var-hap.	max assoc R^2 w/marker SNP
0.0005	11-16	0.00015-0.0060	0.045-0.13	0.91-1.00
0.001	9-14	0.00010-0.0031	0.019-0.050	0.28-1.00
0.002	8-13	0.00010-0.0020	0.0097-0.031	0.06-0.52
0.005	7-10	0.000088-0.001	0.0045-0.011	0.03-0.16

Table 6: Properties of the simulated causal variants at different levels of selection. Selection is measured as the deficiency in fitness of allele carriers relative to non-carriers

permutation of case/control labels provides a null distribution of the test statistic (15) under which there is no association between IBD at the test location and the case/control status of individuals. Since IBD is on a scale of Mbp, at most 3,000 tests can cover the genome. This results in a multiple testing burden that is several orders of magnitude less than that for SNP based GWAS.

To show that population-based IBD mapping can work we present the details of part of the study undertaken by Browning and Thompson (2012). A coalescent simulation including selection and mutation provided a base population with effective size $N_e = 10^4$, over a 200 kbp region of chromosome, representing a functional gene region. The population was then run forward and, at some later time-point, IBD relative to the base population was scored in descendant individuals. In the example summarized here, the effective size of the recent population was $N_e = 10^5$ and the time-depth of IBD was $G = 25$ generations.

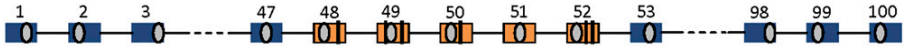


Figure 8: The alternating 1 kbp blocks over a 200 kbp region with the five central blocks containing causal variants. Figure from Browning and Thompson (2012)

For the purposes of association testing, the best SNP in alternating 1 kb blocks was retained, for a total of 100 SNPs (Figure 8). The 5 blocks of the central 10kb (schematically representing the exons of the gene) also contained causal variants that arose in the population simulation. Individuals with ≥ 1 causal variant alleles in the 5 central 1kb blocks are designated as cases with probability 0.1, providing sufficient information for a mapping signal, while still modeling a trait of low penetrance.

Tables 6 and 7 summarize the relevant results from Browning and Thompson (2012). The values in Table 6 show the range of properties of causal variants with different amounts of selection over 100 independent simulations. When selection is weak ($s = 0.0005$), there are somewhat more causal variants, at frequencies up to about 0.5%, but haplotypes carrying causal variants are not rare. These

haplotypes have frequency from 4.5% to 13%, and normally there is a high association between at least one of the causal variants and one of the common SNPs. However, when selection is stronger ($s \geq 0.002$), the frequencies of causal variants are much lower, and the total frequency of haplotypes carrying causal alleles is of the order of 1%. In this case there is rarely a detectable association between any causal variant and any of the 100 common marker SNPs.

selec -tion	# cases= # controls	power assoc.	power IBD	association vs. IBD
0.0005	500	0.87	0.57	assoc.
0.001	500	0.65	0.53	Not-Sig
0.002	1000	0.53	0.87	IBD
0.005	3000	0.47	0.90	IBD

Table 7: Comparison of the power of IBD-based and association tests

The results in Table 7 follow naturally. Here the size of the study is chosen to provide intermediate power values for easier comparison. The association test uses Equation (14) at each of the 100 common marker SNPs and the IBD-based test uses the statistic (15) evaluated at the locations of these common SNPs. When selection is weak ($s = 0.0005$), the association test has higher power. However, when selection is stronger, so that each causal variant has lower frequency, an IBD-based test performs better than an association test. Allelic heterogeneity is a major problem for association testing, unless there is at least one variant with sufficiently high frequency to show association: an IBD-based test is not directly affected by allelic heterogeneity, since each case-case pair has a higher chance of carrying the same causal allele, even though this allele may differ among pairs.

4.4 Model-based mapping likelihoods in populations

In Section 3.2 we saw how IBD $\mathbf{Z}$ could be sampled conditional on marker data $\mathbf{X}$ and a known pedigree structure. In Section 4.2 we saw how these realizations of $\mathbf{Z}$ could be used to compute a likelihood function (13) for use in inferring the locations of DNA underlying trait phenotypes $\mathbf{Y}$. In Section 3.4, we saw how IBD can be realized conditional on marker data in the absence of pedigree information. Finally we now show how these population-based realizations can also be used in genetic mapping. In fact, once the marker data $\mathbf{X}$ have been used to provide realizations of IBD, $\mathbf{Z}$, it is largely irrelevant whether or not they were made conditionally on a known pedigree structure.

For the general trait models considered in Section 4.2 it is usually insufficient to have only pairwise measures of IBD. Even for a single-locus trait model, with hypothesized causal DNA at location j, the probability $\Pr(\mathbf{Y} \mid Z_J; \Gamma_Y)$ (Equation (13)) will depend on the joint IBD state among the individuals observed for the trait. While extension of the methods of Section 2.4 allow efficient com-

putation of this probability for any specified Z_j the number of possible IBD states at a locus is huge (Section 3.4), and effective realization of $\mathbf{Z}$ given $\mathbf{X}$ is a difficult problem. The MCMC methods developed by Moltke *et al.* (2011) and by Zheng *et al.* (2014) are not scalable.

Glazner and Thompson (2015) proposed an alternate approach, in which a joint IBD state among multiple individuals is built up successively from pairwise inferences. The 15-state HMM is run on pairs of individuals as in Brown *et al.* (2012). However, in adding individuals to a joint configuration, the IBD trajectories across the chromosome are constrained by previously realized IBD. Using this method in a simulated example, Glazner and Thompson (2015) showed that joint IBD realized in the absence of an assumed pedigree structure can be used to recover a likelihood across genome locations j using Equation (13). The "gold standards" are likelihoods that would be obtained if IBD were perfectly imputed from marker data $\mathbf{X}$: $\Pr(\mathbf{Y} \mid Z_j; \Gamma_Y)$ at multiple locations j across a chromosome. For IBD realized using haplotypic marker data the approximation is very good, but for genotypic marker data, $\mathbf{X}$, it is less so. Additionally, the method becomes computationally intensive, and performance degrades, as larger sets of related individuals have observed phenotypes, $\mathbf{Y}$. There are additional problems also in the use of Equation (13) for likelihood-based mapping in the absence of a pedigree. First, the usual baseline probability $\Pr(\mathbf{Y}; \Gamma_Y)$ is not available; without a pedigree there is no basis to compute it. Second, no constant baseline works well: the likelihood (13) is affected by the inferred level of IBD across the chromosome and for IBD estimated without the constraints of a pedigree structure, this can vary widely.

Because of unsolved problems in computationally feasible and effective ways to realize joint IBD in the absence of a pedigree structure, we return to the pairwise model of Equation (10) for our final example of IBD-based mapping in the absence of an assumed pedigree. In this model, the local kinships Φ_j are now estimated using the methods of Section 3.4, and the genome-wide kinship Ψ is estimated by averaging the local kinships Φ_j across the genome. This approach was first taken by Day-Williams *et al.* (2011). They used the estimator of locals pairwise kinship Φ_j outlined in Section 3.3 and denoted DW. The resulting estimators, smoothed across the chromosome may additionally be constrained so that each Φ_j for each pair of individuals takes the values $0, 1/4, 1/2$ or 1 (Table 2). By contrast, the joint realizations of IBD among individuals produced by Glazner and Thompson (2015) can be immediately reduced to a set of pairwise realizations of the 15 states of Table 2, and hence to local kinships Φ_j; these estimates are denoted GT. At each j, and for each pair of individuals, the averages across realizations can also be constrained to the values 0, 1/4, 1/2 and 1.

The two estimation methods GT and DW were compared on a simulated example using the variance component log-likelihood-ratio (11) computed at locations j across the chromosome. Since there was no evidence of a genomewide genetic effect, for simplicity it was assumed that $\sigma_a^2 = 0$. The two sets of local estimates were each used constrained and unconstrained. Each of the four log-likelihood curves were compared with the curve that would be obtained if the

actual $\mathbf{Z}$ and hence the realized Φ_j were known at each location j. Details may be found in Glazner and Thompson (2015), but generally the GT estimators performed better than the DW estimators. Also, whereas for GT there was little difference in the results between the constrained and unconstrained versions, for DW the constraint had negative impact especially in regions of high IBD. As noted in Section 3.3, the DW method underestimates IBD particularly in regions of within-individual IBD and multi-gamete IBD.

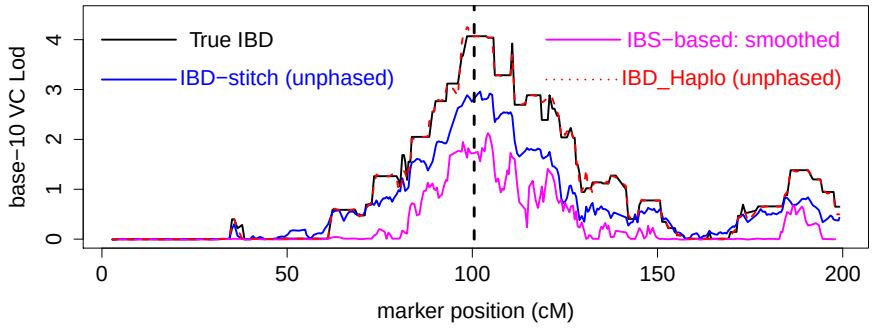


Figure 9: Comparison of DW (magenta), GT (blue), and HMM (red-dashed) estimators of the log-likelihood-ratio (11) on the example of Glazner and Thompson (2015). The black line shows the value that would be obtained is the true realized pairwise IBD were known as each point in the genome.

If a genetic model requires only local pairwise estimates of IBD the HMM method of Brown *et al.* (2012) provides an alternative method. In this case the 15-state HMM is simply run separately on all pairs of observed individuals, to provide estimates of Φ_j across all j. The result of applying this approach to the simulated example of Glazner and Thompson (2015) is shown in Figure 9, together with the results from the the DW and GT approaches. It it seen that the HMM method almost perfectly recovers the actual realized pairwise IBD in this example. Given the difficulties of estimation of IBD among multiple individuals, the use of models and methods that require only pairwise estimates provide an attractive alternative.

In Section 4.2 we showed how, on a defined pedigree, realizations of $\mathbf{Z}$ given genetic marker data $\mathbf{X}$ could be used to provide estimates of linkage likelihoods $\Pr(\mathbf{Y} \mid \mathbf{X}, \Gamma)$ for locations $\lambda = \{j\}$ across a chromosome. In this section we have shown how the same may be accomplished in populations, using a population prior model for IBD (Section 2.2). Once realizations of $\mathbf{Z}$, jointly among individuals and across a chromosome, are obtained conditional on genetic marker data $\mathbf{X}$ there is no essential difference whether these were made with or without the assumption of a pedigree structure. The pedigree structure provides a more informative, and sometimes overly constraining, prior, but modern SNP data at

multiple markers can compensate for the lack of pedigree information.

This raises the attractive possibility of combination of pedigree and population-based IBD. The pedigrees of any family study exist within a population, and founders within and between pedigrees may be related. Methods to combine IBD inferred within pedigrees with that inferred among founder members have been implemented (Saad *et al.*, 2016), but there are several issues. First founders of pedigree structures are often unobserved: populations-based inference of IBD is practical only for individuals with fully observed genotypic data. Second, there are multiple individuals who may be related outside the defined pedigree structures. Population-based realizations of IBD jointly among multiple individuals is intractable, and multiple pairwise estimates not easy to process successfully. There remains an additional issue. Within a pedigree, the maternal/paternal origins of haplotypes can be realized where there are informative data. In other cases, for example for the two haplotypes of founders, there is no information (even with data) on which haplotype is maternal and which paternal, but this is irrelevant to within-pedigree IBD. In the population-based context, even if the IBD inference implies fully correct phasing, it is likewise arbitrary which haplotype is designated the maternal/paternal one of the individual. Combining population and pedigree IBD faces the intractable challenge of resolving the multiple pairings of each individual's two haploid genomes with the labelling in each population-based realization of IBD. Thus while IBD provides a natural unifying framework in which to combine pedigree- and population-based inferences, there remain challenges for successful implementation of methods.

5 Summary

We have shown how IBD $\mathbf{Z}$ can be inferred from genetic marker data $\mathbf{X}$, and then used to provide evidence for genome locations at which the DNA variants may be causal for trait phenotypes $\mathbf{Y}$. Rather than considering directly the association between $\mathbf{X}$ and $\mathbf{Y}$, we consider this association through the lens of descent $\mathbf{Z}$. In fact, a basic assumption of our models is that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$. In the three main sections of the text we have considered: first, probability models for $\mathbf{Z}$; second, inference of $\mathbf{Z}$ from$\mathbf{X}$; and third, use of this inferred or realized $\mathbf{Z}$ to map DNA underlying $\mathbf{Y}$

In Section 2 the focus was on the models for IBD, or $\mathbf{Z}$, and then on data $\mathbf{X}$ or $\mathbf{Y}$ given $\mathbf{Z}$. We considered probability models for $\mathbf{Z}$, both in defined pedigrees and among members of a population. It is important to consider not only IBD at separate locations, but also how it changes across a chromosome. Because DNA descends generation to generation is large segments, even remote relatives will (if they share any genome IBD) share segments that are of order of millions of base pairs long, and will contain many SNP markers. Also in Section 2 we considered probabilities of marker data and trait data conditionally on $\mathbf{Z}$. They key point here is that, for these probabilities, it is irrelevant whether or not the pedigree structure is known.

In Section 3 we turn to the inference of $\mathbf{Z}$ given marker data $\mathbf{X}$. We first

consider the case of defined relatives, where the pedigree structure provides a strong, but sometimes overly-constraining, prior. Usually not all members of a pedigrees are typed, but pedigrees can only be validated among current individuals for whom marker genotypes are available. In some studies also there may be issues of genotypic error; for computational reasons, pedigree-based methods assume that marker data are observed without error. We then turn to the inference of IBD in population, the purpose of the IBD model being to provide a flexible and tractable prior for inference. One important aspect of this flexibility is that is allows for error in the observation of marker genotypes. SNP typing is quite accurate, but there are very large numbers of SNPs. We consider first genome-wide measures of IBD. We point out issues with methods that treat all SNPs equally, and take no account of their dependence due either to allelic association (LD) or to their physical locations. We describe one method of adjusting for LD, but our major focus again is on the segmental natural of DNA descent. Since individual SNPs are very uninformative, combining multiple SNPs in detecting IBD segments is of key importance. Using a model for the dependence of descent across SNP markers has two important consequences. First estimates of genome-wide IBD proportions are greatly improved. Second, and essential for gene mapping purposes, $\mathbf{Z}$ is realized at locations across the genome: the actual locations of segments of IBD are detected.

Finally, in Section 4 we show how $\mathbf{Z}$ inferred from marker data $\mathbf{X}$ can be used to map DNA that is causal to trait data $\mathbf{Y}$ against the genetic marker map. Again we consider first the case of pairs or groups of individuals whose pedigree relationships are know. This includes approaches such as that of affected relative pairs for binary traits, and variance-component models for mapping quantitative trait loci (QTL). We then extend to more general models for phenotypes $\mathbf{Y}$ and show how realizations of $\mathbf{Z}$ conditional on $\mathbf{X}$ can be used to obtain Monte Carlo estimates of linkage likelihoods $\Pr(\mathbf{Y}|\mathbf{X})$ for a specified trait model, and specified hypotheses of the location(s) of causal DNA. Next we return to populations, and consider an IBD-based analogue of case-control studies, showing that where different rare variants in a functional gene can cause the trait, the IBD-based approach outperforms a GWAS test. IBD-based tests can address allelic heterogeneity both in pedigrees and in populations. Finally we return to linkage likelihoods, on the basis of IBD inferred in populations where pedigree relationships are unknown. We consider the complexities of multi-individual IBD and suggest that often a variance-component model that requires only pairwise IBD may be ore useful. However, the more fundamental message is that it is largely irrelevant to subsequent analysis whether IBD is inferred under a population model or on a defined pedigree. All pedigrees exist within a broader population framework, the IBD framework permits the combination of population and pedigree information.

References

Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon, 2002 Merlin

– rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics **30**: 97–101.

Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics **62**: 1198–1211.

Blangero, J., J. T. Williams, and L. Almasy, 2000 Robust LOD scores for variance component-based linkage analysis. Genetic Epidemiology **19**: S8–S14. Suppl. 1.

Brown, M. D., C. G. Glazner, C.Zheng, and E. A. Thompson, 2012 Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics **190**: 1447–1460.

Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. American Journal of Human Genetics **86**: 526–539.

Browning, S. R., and E. A. Thompson, 2012 Detecting rare variant associations by identity by descent mapping in case-control studies. Genetics **190**: 1521–1531.

Chapman, N. H., A. Q. N. Jr., R. Bernier, K. Ankeman, H. Sohi, J. Munson, A. Patowary, M. Archer, E. M. Blue, S. J. Webb, H. Coon, W. H. Raskind, Z. Brkanac, and E. M. Wijsman, 2015 Whole exome sequencing in extended families with autism spectrum disorder implicates four candidate genes. Human Genetics **134**: 1055–1068.

Cox, D. R., 1962 *Renewal Theory*. Methuen and Co., London, UK.

Day-Williams, A. G., J. Blangero, T. D. Dyer, K. Lange, and E. M. Sobel, 2011 Linkage analysis without defined pedigrees. Genetic Epidemiology **35**: 360–370.

Donnelly, K. P., 1983 The probability that related individuals share some section of genome identical by descent. Theoretical Population Biology **23**: 34–63.

Elston, R. C., and J. Stewart, 1971 A general model for the analysis of pedigree data. Human Heredity **21**: 523–542.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87–112.

Glazner, C. G., and E. A. Thompson, 2015 Pedigree-free descent-based gene mapping from population samples. Human Heredity **80**: 21–35.

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics **8**: 229–309.

Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics **2**: 3–19.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genome Research **91**: 47–60.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetical Research Cambridge **93**: 47–64.

Karigl, G., 1981 A recursive algorithm for the calculation of gene identity coefficients. Annals of Human Genetics **45**: 299–305.

Koepke, H. A., and E. A. Thompson, 2013 Efficient testing operations on dynamic graph structures using strong hash functions. Journal of Computational Biology **20**: 551–570.

Lander, E. S., and D. Botstein, 1987 Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science **236**: 1567–1570.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences (USA) **84(8)**: 2363–2367.

Lange, K., and E. Sobel, 1991 A random walk method for computing genetic location scores. American Journal of Human Genetics **49**: 1320–1334.

Lauritzen, S. J., 1992 Propagation of probabilities, means and variances in mixed graphical association models. Journal of the American Statistical Association **87**: 1098–1108.

Leutenegger, A., B. Prum, E. Genin, C. Verny, F. Clerget-Darpoux, and E. A. Thompson, 2003 Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics **73**: 516–523.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature **475**: 493–496.

McPeek, M. S., 1999 Optimal allele-sharing statistics for genetic mapping using affected relatives. Genetic Epidemiology **16**: 225–249.

Mendel, G., 1866 Experiments in Plant Hybridisation in *English translation and commentary by R. A. Fisher*, edited by J. H. Bennett. Oliver and Boyd, Edinburgh, 1965.

Moltke, I., A. Albrechtsen, T. Hansen, F. C. Nielsen, and R. Nielsen, 2011 A method for detecting IBD regions simultaneously in multiple individuals — with applications to disease genetics. Genome Research **21**: 1168–1180.

Morton, N. E., 1955 Sequential tests for the detection of linkage. American Journal of Human Genetics **7**: 277–318.

Peter, B., E. M. Wijsman, A. Q. N. Jr., M. Matsushita, K. L. Chapman, I. B. Stanaway, J. Wolff, K. Oda, V. B. Gabo, and W. H. Raskind, 2016 Genetic candidate variants in two multigenerational families with childhood apraxia of speech. PLOS ONE: e0153864.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**: 904–909.

Saad, M., A. Q. Nato, F. L. Grimson, S. M. Leweis, L. Brown, E. M. Blue, Thornton, E. A. Thompson, and E. M. Wijsman, 2016 Identity-by-descent estimation with population- and pedigree-based imputation in admixed family data. BMC Proceedings **10(Suppl 7)**: 295–301.

Smith, C. A. B., 1953 Detection of linkage in human genetics. Journal of the Royal Statistical Society, B **15**: 153–192.

Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics **58**: 1323–1337.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics **91**: 1011–1021.

Su, M., and E. A. Thompson, 2012 Computationally efficient multipoint linkage analysis on extended pedigrees for trait models with two contributing major loci. Genetic Epidemiology **38**: 602–611.

Suarez, B. K., J. Rice, and T. Reich, 1978 The generalized sib pair IBD distribution: Its use in the detection of linkage. Annals of Human Genetics **42**: 87–94.

Tavaré, S., and W. J. Ewens, 1997 The multivariate Ewens distribution, pp. 232–246 in *Discrete Multivariate Distributions*. Wiley, New York, NY.

Thompson, E. A., 2000 *Statistical Inferences from Genetic Data on Pedigrees*, Volume 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Beachwood, OH.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Wang, B., S. Sverdlov, and E. A. Thompson, 2017 Efficient estimation of realized kinship. Genetics **205**: 1063–1078.

Weeks, D. E., and K. Lange, 1988 The affected pedigree member method of linkage analysis. American Journal of Human Genetics **42**: 315–326.

Zheng, C., M. K. Kuhner, and E. A. Thompson, 2014 Joint inference of identity by descent along multiple chromosomes from population samples. Journal of Computational Biology **21**: 185–200.