1 Introduction

Every child is born with a few *de novo* mutations, DNA sites where they differ from their parents and from most other humans. Most of the variants created this way die out within a few generations, but a minority of them spread to hundreds or thousands of the child's descendants and contribute to widespread human genetic variation [1, 27]. By mathematically modeling the emergence and spread of new alleles, population geneticists can make inferences about ancient periods of growth, decline, interbreeding, and the emergence of modern ethnic groups, as well as discover links between genetic and phenotypic variation.

Given DNA from one individual, it is much cheaper to genotype a few thousand genetic loci than to ascertain the entire genome sequence, so companies like Illumina and Affymetrix manufacture single nucleotide polymorphism (SNP) chips that can selectively ascertain the states of between 10,000 and 1,000,000 of the most variable sites in humans. By focusing on fewer genetic sites, one can afford to genotype those sites in more individuals, and this approach has been used since the invention of pedigree analysis to find many sites in the genome that correlate with disease risk or recent positive selection [36]. A problem with SNP chips, however, is that they omit sites where variant alleles arose too recently to spread to a significant fraction of the human population. Although none of the three billion sites that are omitted from a SNP chip is especially variable on its own, together they harbor a vast amount of additional genetic information [1, 26]. This hard-to-detect variation is a clear candidate harbor for "missing heritability" in disease genetics, where known genetic risk factors usually fail to account for the full heritability of complex diseases [29, 51].

One way to detect more low-frequency variants will be to gather more genotype and sequence data, working to make this process cheaper through improvements in biotechnology. Another approach, however, is to extract more information from available data sets by modeling a process known as *linkage disequilibrium* (LD). Even when site X does not appear on a chip being used to to gather data, it can still be possible to infer that two sequences match at site

X by looking for matching at sites close to X. DNA is passed from parents to children in continuous blocks between recombination sites; when two sequences share a rare allele, it is likely that the allele was inherited from a recent common ancestor along with a block of surrounding DNA containing some sites that appear on the SNP chip [10, 31].

Linkage disequilibrium affects the distribution of heterozygous sites (*hets*) in every diploid genome, even in outbred populations. If every site in the genome had an independent probability m of being a het, then the probability of an L-base region being devoid of hets, or *identical by state* (IBS) would be $(1-m)^L \approx e^{-L}$. The frequency of L-base regions of homozygosity (ROHs) is not observed to decline exponentially with L, however [41, 55], and the excess of long ROHs can be accounted for by modeling LD. If, for example, an individual's parents are ninth-degree cousins, there is only a one-in-2^{20} chance that both alleles at a given site in the child's DNA were inherited from the parents' most recent common ancestor, but given that both alleles *were* both inherited from that ancestor, the child is likely to be homozygous over 10 megabases of surrounding DNA [31]. Ten generations is not enough time for meiosis to break the DNA into smaller heritable pieces, and in general, the length of a homozygous stretch is inversely proportional to the age of the ancestor that the matching haplotypes derive from.

The key to understanding how hets are placed is understanding how coalescence time, or time to common ancestry, varies from site to site across the genome. We define *ancestral recombination sites* (ARs) to be loci where two neighboring allele pairs coalesce at different times, and say that an alignment is *identical by descent* (IBD) if it has no interior ARs (See Figure 1 for an illustration of IBD vs. IBS).

A consequence of coalescent theory is that hets are placed randomly within an IBD region, with their density proportional to the region's coalescence time t ($t = 0$ being the present and larger t's being more ancient). As we move from left to right across a region of IBD, each base has a constant probability of being a het and a constant probability of being an AR and ending the IBD stretch.

In human DNA, the het probability μt is about 2.5 times the AR probability ρt, such that each IBD region contains about 2.5 total hets. The length of the region will vary inversely with t, however, making the local density of hets very small when t is very small.

Commercial chips with at most 1,000,000 SNP sites can detect at most 10% of the hets in a diploid sequence. This suggests that, on average, an IBD region will contain 0.25 hets that are detectible with a 1,000,000 chip and that $e^{-0.25} > 0.77$ of all maximal IBD regions will appear IBS based on genotype data. In contrast, only $e^{-2.5} \approx 0.082$ of maximal IBD regions will appear IBS based on sequence data.

Although definitions of IBD differ widely in the literature, IBD between two sequences is usually taken to imply IBS at the level of genotype data, and we do not intend to create confusion by defining IBD such that it does not imply IBS. Rather, we note that sequence-level IBS will only be true of about $0.082/0.77 \approx 0.11$ of the regions that are inferred to be IBD by a program like BEAGLE, which used IBS at the genotype level to find segments of shared ancestry. In contrast, 77% of the 1 MB regions that we call IBD should also be identified as IBD by BEAGLE [10]. When looking for IBD in sequence data, it seems useful to drop the assumption that IBD implies IBS, just as it was necessary to change the definition of IBD when moving from pedigree analysis to the study of unrelated individuals.

The terms IBD and IBS were in fact both coined in the context of pedigree analysis, where a family with a history of a disease phenotype is scrutinized for genetic variants that might contribute to the appearance of that phenotype. Related individuals are genotyped at a sparse set of markers, and those markers are used, together with the family relationship pedigree, to find haplotypes that were often transmitted from diseased ancestors to diseased offspring [34, 37]. IBD sharing makes it likely that two individuals match at a long stretch of unobserved DNA, and inferring this matching is essential given that the variants causing the disease will almost certainly not be among the few directly genotyped marker sites.

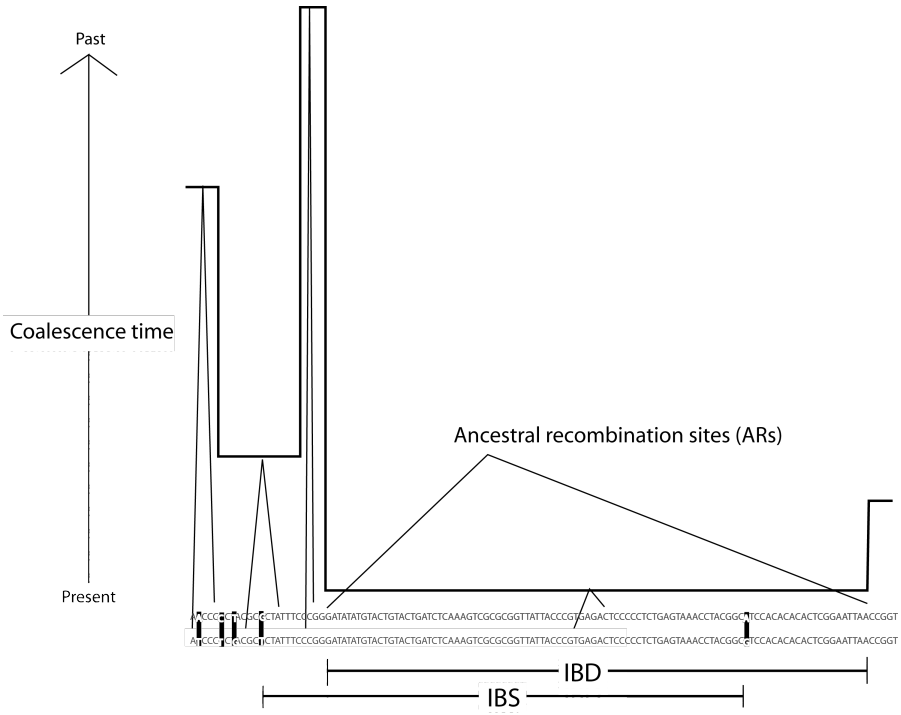

Figure 1: This picture illustrates the difference between IBS and IBD. IBS depends only on the observable differences between two sequences, while IBD depends on their hidden history: how long ago each site coalesces. Two sequences are IBD if each base coalesces at the same time, and IBS if each base matches by state (there are no internal hets). Long regions of IBS usually overlap with long regions of IBD, but as shown here, the regions rarely coincide exactly.

More data is available in a genome-wide association study (GWAS), where thousands of cases and controls are typed at hundreds of thousands of markers. However, it is impossible to know the family relationships among so many study individuals, making direct IBD inference more difficult than in a linkage study, and it is still likely that the causal variants will not be directly genotyped.

Rather than working to infer the genealogies of unobserved stretches of DNA, GWASs regard typed markers as one-to-one proxies for untyped markers, working to construct genotype sets for which each unobserved allele is usually coinherited with an observed allele. If the presence of allele A at observed locus x means that there is a 90% chance of observing allele B at locus y, then even if B is causal and A is not, it may be possible to observe a correlation between the presence of A and the disease. A strong pairwise association between A and B translates to a high correlation coefficient $r^2(A, B)$, which is calculated from the allele frequencies $f_{A(x)}$ and $f_{B(y)}$ along with the haplotype frequency $f_{A(x)B(y)}$:

$$r^2(A, B) = \frac{(f_{A(x)B(y)} - f_{A(x)}f_{B(y)})^2}{f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})} \tag{1}$$

A standard measure of a genetic tag set's efficacy is the percentage of untyped variable sites that are within $r^2 \geq 0.8$ of a typed SNP (see e.g. [6]).

By the definition given above, $r^2(A, B)$ is a statement about how often A and B occur together in extant individuals, not a statement about how much history the alleles have in common. McVean showed that $r^2(A(x), B(y))$ is related to the covariance between the coalescence times at sites x and y [42]; IBD histories are more probable when r^2 is close to 1, but knowing $r^2(A, B)$ is not sufficient to know the likelihood of IBD in the stretch between x and y. Similarly, Hayes, et al. showed that the mean r^2 for markers L bases apart is close to the frequency of L-base IBD stretches in the genome, but that their measure of IBD sharing has a lower variance than r^2 does, capturing strictly more information about the hidden history of the sequences [17].

Some have claimed that pairwise r^2 values behave badly when input into multivariate GWAS analyses, and that measures of IBD probability behave

much better. Terwilliger and Hiekkalinna argue that it is dangerous to assume that the correlation between a tag and a variant will be statistically independent of the correlation between the variant and the phenotype, and that this is a fatal flaw in the paradigm of using tag SNPs as one-to-one proxies for unobserved SNPs. In contrast, they argue that IBD sharing *should* be independent of whether any loci involved are functional [56], and that linkage studies may be inherently more powerful than GWAS as a result. Whether or not they are correct, the best of both worlds solution may be to conduct GWAS as much like linkage studies as possible, finding ways to look for IBD sharing, rather than simple IBS association, in genetic data sets that have no accompanying pedigree data.

Imputation can be viewed as a step toward making GWAS more like linkage studies, inferring IBD with the help of population genetics rather than pedigrees [38]. To avoid assuming that the effects of untyped variants will automatically show up by proxy association, these variants are imputed into test sequences and screened for association directly. Imputation is performed where IBD sharing is suspected between a sample and a reference haplotype, taking advantage of the good evidence for IBD that is provided by long IBS marker strings. We are able to compute precisely how long these marker strings must be for p(IBD|IBS) to be sufficiently close to 1, analytically predicting when imputation should be reliable.

Detecting IBD is especially important when causal variants are very rare or have very modest phenotypic effects. Several variants that affect the same condition may be clustered around an important protein or promoter, in which case it may be possible to pool their signals, i.e. regard the whole region as single locus where haplotypes are the alleles. The number of individuals with causal variants in the region should exceed the number with variants at any particular locus, and the pooled signal of these variants may just reach the threshold of detectability [9, 51, 53]. However, this approach depends on the ability to tell haplotypes apart based on marker IBS, and we will show that inferring a 1000-base haplotype with 99% accuracy requires imputing from nearly a megabase

of standard IBS marker data.

The probability p(IBD|IBS) depends on population history in a complex way; although long IBD tracts are most common in DNA from inbred groups, inbreeding actually increases the probability that a long IBS tract is not IBD [37, 52, 55]. False discoveries abound in linkage studies that do not adequately account for hidden founder relatedness, particularly with regard to long genetic loops that are seldom recorded in pedigrees [34, 37, 52]; however, the dependence of IBS sharing on population history can be useful as well as confounding, since the length distribution of shared IBS contains more information about population substructure than simpler measures like the coefficient of relatedness. Jakkula, et al., for example, found that the Finnish sub-populations have similar inbreeding coefficient distributions but differ significantly in their patterns of homozygosity and IBS sharing [28]. Similarly, Kong, et al. found long IBS sharing to be common in Iceland, though the average inbreeding coefficient (2.5×10^{-4}) was not especially high. In a collection of 35,528 Icelanders who were genotyped for a particular 10 Mb region, all but 1,995 shared that region IBS with another genotyped individual who was not closely related to them, enough to allow for long-range phasing within the population at large [31].

There exist several algorithms for estimating p(IBD|IBS), some conditioning on haplotype frequencies and some only on inheritance models. The data-dependent algorithms have the advantange of specificity, but they consistently underestimate p(IBD|IBS) because of the way they incorporate their test haplotype into their prior [9, 33, 46]. They can confirm that a medically interesting region is likely to be IBD, but are less useful for using IBD to study population history. p(IBD|IBS) has not been computed exactly with respect to the neutral coalescent, and we believe we are the first to compute it with respect to the sequentially Markovian coalescent [43].

Previous methods for estimating p(IBD|IBS) that do not condition on allele frequencies have begun to deduce the impact of history on genome-wide patterns of IBD [11, 17, 54, 55, 57]. However, most of them make assumptions that break down at certain segment lengths and marker densities, which prevents

them from making use of all available marker information. The PLINK hidden Markov model, for example, will only calculate $p(\text{IBD}|\text{IBS})$ between markers that are in linkage equilibrium with each other [33, 51]; their precision is limited by the sparseness of unlinked marker sets. A related assumption, which is implicitly made in all of the literature we found, is that the lengths of adjacent IBD segments are independently distributed [9, 17, 46, 51, 55], and we will show in Section 6 how this breaks down for large, dense data sets. Our method, in contrast, captures the dependence between the lengths of neighboring IBD segments, and can assume arbitrarily dense marker data without losing any accuracy. Given inputs of population size history, mutation rate, and recombination rate, we predict an ROH distribution that can be verified in genome data. After adjusting for the presence of sequencing errors, we are able to accurately predict the distribution of ROHs found in eleven complete human genome sequences.

Given that sequencing is much more costly than genotyping, we also adjust our method to predict IBS given a thinned-down set of markers. Our theory correctly predicts the distribution of segments that appear homozygous based on incomplete knowledge of the hets in the genome data, quantifying the correlation between IBS at the genotype level and IBS at the level of the complete sequence.

We also extend our theory to the case of unphased diploid sequences, deviating from the SMC slightly but checking the results against a full coalescent simulation. When phasing ambiguities are accounted for in this way, $p(\text{IBD}|\text{IBS})$ can be used to estimate the accuracy of an attempt at imputation and/or haplotype resolution. We conclude that both efforts become much more accurate if the ends of an IBS alignment are not considered likely to be IBD; when IBS is measured in a way that detects a het every 10,000 bases, it seems prudent to discard 10^5 bases from each end of an alignment, after which the probability of IBD is as great as if the full sequences were known. Finally, we estimate the accuracy spread of the imputation calls made from a panel of n reference haplotypes, showing that a thousand references should be sufficient in a population where recent exponential growth has not broken up moderately long stretches

of IBD sharing.

2 Computing the probability of identity by state

Since

$$p(\text{IBD}|\text{IBS}) = \frac{p(\text{IBD\&IBS})}{p(\text{IBS})}, \tag{2}$$

where p(IBD&IBS) is easy to compute (see equation (2.1)), the crux of our approach will be calculating p(IBS) given sequence length and the history of the effective population size. In section 2.1, we treat the case of constant effective population size, while section 2.3 describes how to condition on any locally constant population size history.

2.1 Constant effective population size

Let L be the length of an alignment between two haplotypes sampled at random from a diploid population of effective size N. Assume that the DNA undergoes m mutations per base per generation and r recombinations per base per generation, letting $\mu = 4Nm$ and $\rho = 4Nr$. We will hereafter measure time in units of $2N$ generations.

The alignment will coalesce at time t, both IBD and IBS, if and only if the following events coincide:

1. The leftmost locus coalesces at time t without mutating (probability $e^{-t(1+\mu)}dt$)
2. No other base in either sequence undergoes a mutation or a recombination between time zero and time t (probability $e^{-t(L-1)(\mu+\rho)}$)

From this observation, it follows that

$$p(\text{IBD\&IBS}) = \int_{t=0}^{\infty} e^{-t(L-1)(\mu+\rho)} \cdot e^{-t(1+\mu)}dt = \frac{1}{1 + L\mu + (L-1)\rho}. \tag{3}$$

In an analogous way, we will derive the probability $p_L(\text{IBS}|t)dt$ that the alignment coalesces IBS with its rightmost base coalescing at time t. We proceed