Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies

Sharon R. Browning*,[1] and Elizabeth A. Thompson†
*Department of Biostatistics, University of Washington, Seattle, Washington 98195
and †Department of Statistics, University of Washington, Seattle, Washington 98195

ABSTRACT Identity-by-descent (IBD) mapping tests whether cases share more segments of IBD around a putative causal variant than do controls. These segments of IBD can be accurately detected from genome-wide SNP data. We investigate the power of IBD mapping relative to that of SNP association testing for genome-wide case-control SNP data. Our focus is particularly on rare variants, as these tend to be more recent and hence more likely to have recent shared ancestry. We simulate data from both large and small populations and find that the relative performance of IBD mapping and SNP association testing depends on population demographic history and the strength of selection against causal variants. We also present an IBD mapping analysis of a type 1 diabetes data set. In those data we find that we can detect association only with the HLA region using IBD mapping. Overall, our results suggest that IBD mapping may have higher power than association analysis of SNP data when multiple rare causal variants are clustered within a gene. However, for outbred populations, very large sample sizes may be required for genome-wide significance unless the causal variants have strong effects.

THE idea of using identity-by-descent (IBD) haplotype sharing to detect signals of disease-causing variants in population samples is not new (Houwen *et al.* 1994; Te Meerman *et al.* 1995); however, the greatly increased density of SNP markers now makes it possible to detect much smaller segments of IBD. New statistical methods for detecting such IBD have been proposed (Purcell *et al.* 2007; Kong *et al.* 2008; Thomas *et al.* 2008; Leibon *et al.* 2008; Gusev *et al.* 2009; Albrechtsen *et al.* 2009; Thompson 2009; Bercovici *et al.* 2010; Browning and Browning, 2010, 2011; Han and Abney 2011; Brown *et al.* 2012), and it is possible to determine pairwise IBD sharing in a large sample over the whole genome to a resolution of approximately 2 cM (Browning and Browning 2011). In this article we investigate the power of IBD mapping to detect associations for complex diseases and compare this with the power of SNP association mapping.

Throughout this article, IBD is genomic-location specific. That is, we have information from genetic data on whether two individuals share alleles identical by descent at a certain genomic position. We focus on segments of IBD that are due to recent shared ancestry. For example, IBD segments from shared ancestry 25 generations ago have average length of 2 cM.

Two classes of statistics have been proposed for IBD mapping. The first, which we call "pairwise" statistics, use IBD detected between pairs of individuals. The rate of IBD in case/case pairs is compared to the rate of IBD in either control/control pairs or non-case/case pairs (control/control and control/case pairs) (Purcell *et al.* 2007). The pairwise approach is somewhat similar to affected relative pair linkage analysis, with important differences. Affected relative pair linkage analysis involves pairwise IBD, but only within relative pairs, not across relative pairs. Also, control individuals are not needed in affected relative pair linkage because the background rate of IBD (the rate of IBD in pairs of unaffected relatives of the same type) is assumed to be known. In IBD mapping, the exact degree of relationship is unknown, and the rate of detected IBD tends to vary along the genome due to differences in informativeness and stochastic differences in haplotype genealogies. The use of control individuals allows the analysis to account for these differences in background rates of IBD.

doi: 10.1534/genetics.111.136937
Manuscript received November 16, 2011; accepted for publication January 17, 2012
Available freely online through the author-supported open access option.
[1]Corresponding author: Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195. E-mail: sguy@uw.edu

The second class of statistics, which we call "clustering" statistics, cluster haplotypes into IBD classes at a locus. All haplotypes within a class are IBD with each other at the locus. An individual is a member of two IBD classes at a locus because the individual has two haplotypes, although the two classes will be the same if the individual is homozygous by descent. Clusters are tested for association with case-control status (Gusev *et al.* 2011). The clustering approach is difficult because IBD is not usually estimated with 100% certainty or 100% power. Thus one must determine how to resolve inconsistencies, such as when haplotypes A and B are estimated to be IBD, and B and C are estimated to be IBD, but A and C are not estimated to be IBD. Also, for IBD due to recent ancestry the IBD clusters will be very small in an outbred population, so that testing individual clusters will tend to have low power.

Whichever statistic is used, IBD mapping focuses on signals from rare variants. That is because coalescence times will be shorter for haplotypes carrying the same recent (hence rare) variant, resulting in larger segments of IBD that are more detectable. In small populations, all coalescence times are relatively short, so the variants need not be as rare. Various methods are available for detecting IBD segments. Resolution of the methods differs, but generally power is low for detecting segments shorter than 2 cM with current genome-wide SNP panels in Europeans (Browning and Browning 2011).

IBD mapping is performed on data that are also suitable for standard association analysis. If the data are sequence data, with all rare variants genotyped, then the data contain perfect information about IBD for the purposes of determining association. Variants that are identical can be assumed to be IBD; even if there is recurrent mutation and the identical variants are not IBD, it is only the identity of the variant that matters when directly testing a putative causal variant. Hence detected IBD segments cannot add further information. Thus, for sequence data, standard single-variant association testing should be more powerful than IBD mapping. From the point of view of merely comparing testing procedures, standard single-variant association testing pays the cost of higher multiple testing correction, but IBD mapping suffers from incomplete information (not all IBD is detected) and added noise. For the pairwise IBD statistics, the background rate of IBD provides noise; for the clustering statistics, IBD clusters that do not represent actual variants provide noise. The pairwise statistics allow aggregation from several causal variants located proximally (*e.g.*, within the same gene), which can add power. With sequence data, specialized association tests can aggregate over multiple rare variants and have the advantage that functional information, such as whether a variant is a missense mutation, can be utilized (Li and Leal 2008; Madsen and Browning 2009; Wu *et al.* 2011).

In genome-wide SNP array data, many rare variants are not typed, allowing the possibility of a power advantage for IBD mapping. For non-IBD-based testing, both single-SNP and haplotypic tests are possible. Haplotypic tests can be useful because haplotypes can tag rare variants that have not been genotyped. The line between IBD mapping and haplotypic testing with long haplotypes is somewhat blurred. The IBD clustering tests are similar to haplotype clustering tests (Browning 2006), while the IBD pairwise tests are similar to haplotype sharing tests (Van Der Meulen and Te Meerman 1997). Although haplotypic association tests have theoretical advantages over SNP association tests for detecting associations with rare variants, several practical issues have limited their usefulness. First, haplotypic tests are very susceptible to differential genotype error between cases and controls that can create apparently unusual haplotypes primarily carried by either cases or controls (Browning and Browning 2008). IBD mapping is less susceptible to this problem because genotype errors generally can make it more difficult to detect IBD but would not usually induce false-positive IBD. Thus a small rate of genotype error at a locus, even if differentially distributed between cases and controls, will not be likely to create false-positive IBD mapping signals. Second, haplotypic tests (and IBD mapping tests) are focused on the effects of rare variants. However, power is low to detect association with a rare variant unless it has a very strong effect. Although aggregation over multiple rare variants is possible for haplotypic tests (Zhu *et al.* 2010), and may add power, the effectiveness of the aggregation is reduced because of the lack of functional information about the rare variants being tagged by the haplotypes. Pairwise IBD tests also aggregate over variants, and one advantage of this approach is that one does not need to decide which variants to aggregate, as the aggregation is done automatically by the test.

Some causal rare variants may have large effect sizes relative to the typically small effect sizes seen for common variants in complex traits (Manolio *et al.* 2009), as rare variants tend to be recent and thus have had less time to be removed from the population by natural selection. If the rare variant effects are very large (such as in fully penetrant Mendelian diseases), these effects would have been found through previous family-based linkage studies. However, there may be a middle ground in which multiple rare variants of moderate effect size play a key role in the etiology of some diseases. Such situations may be ideal for IBD mapping.

Few published studies have attempted IBD mapping in unrelated samples. Albrechtsen *et al.* (2009) and Moltke *et al.* (2011) analyzed small numbers (five to seven) of unrelated, mostly Danish, breast and ovarian cancer patients who were known to carry the same *BRCA1* mutation and showed that these individuals were detectably IBD around the gene. Gusev *et al.* (2011) applied DASH, a haplotype-testing/IBD mapping program, to data from samples from the island of Kosrae (a small founder population) and to case-control data from the UK, and found known associations and some novel associations, including one in the Kosrae data that replicated in a European cohort. Francks *et al.* (2010) used PLINK's IBD mapping method on two cohorts of

400–500 schizophrenia cases and similar numbers of controls and obtained a *P*-value of 1.6×10^{-5} on chromosome 19q. By adding approximately 1000 familial bipolar cases and a similar number of controls they were able to strengthen the signal to $P = 2.6 \times 10^{-6}$. It remains to be seen whether this signal will replicate in other data sets.

Methods

Pairwise IBD test

Our IBD test compares rates of IBD in case/case pairs of individuals and non-case/case pairs of individuals. At each position we calculate the rate for each of the two groups (the group of case/case pairs and the group of non-case/case pairs) and subtract off the genomic average for each group; we then take the difference between the two groups. We perform a one-sided test: that is, we test for case/case pairs having a higher adjusted rate of IBD than non-case/case pairs. Because IBD segments are not independent between pairs, we use permutation of case-control labels to assess significance. This testing procedure is the same as in Purcell *et al.* (2007). Python programs implementing the IBD test that we used for the WTCCC type 1 diabetes data can be downloaded from http://faculty.washington.edu/sguy/ibdmapping.html.

SNP association test

For comparison with the pairwise IBD test, we consider allelic association tests. The power of these tests is reduced if there is allelic heterogeneity (multiple causal mutations in a gene), because each single-marker test is likely to tag at most one of the causal variants. Power also tends to be low for rare causal variants, because these are not well tagged by genome-wide SNP panels.

Genome-wide association analyses involve hundreds of thousands of association tests. Thus, multiple-testing correction is required to avoid obtaining large numbers of false-positive results. The multiple-testing correction depends on the panel of SNPs used for testing as well as on the population from which the sample is derived. Gao *et al.* (2010) find that a nominal *P*-value of 1.4×10^{-7} corresponds to an adjusted *P*-value of 0.05 for Illumina 1 M data in European American data while the corresponding nominal *P*-value threshold for Affymetrix 500K data are 2.5×10^{-7}. Dudbridge and Gusnanto (2008) extrapolate from WTCCC (UK) data to a *P*-value threshold of 7.2×10^{-8} for infinitely dense SNP data with allele ascertainment matching that of the Affymetrix 500K data. Pe'er *et al.* (2008) estimate a multiple testing threshold for all common SNPs (frequency >5%) of around 10^{-7} in Europeans and 5×10^{-8} in Africans. Recently, a threshold of 5×10^{-8} has become a *de facto* standard regardless of the populations being studied (*e.g.*, Lowe *et al.* 2009).

The effective population size to consider for SNP association testing is different than that for IBD testing, as the relevant time interval differs. IBD testing concerns mostly the effective population size in the past G generations, where $G = 25$, for example (the recent effective population size), whereas SNP association testing is concerned with much larger timescales (the long-term effective population size). For a population of changing size, the relevant effective population size is close to the minimum (effective) population size over the time period of interest (Wright 1931). Thus population bottlenecks, for example during the out-of-Africa migration(s), can have a large impact on long-term effective population size, and hence on the multiple testing correction for SNP association tests.

Results

Simulation study

We performed a simulation study to compare standard SNP association testing with pairwise IBD testing. We used SFS_CODE (Hernandez 2008) to generate sequence data from a base population of constant size $N = 10{,}000$. The population size influences the scaled rates of mutation, $4N\mu$, and selection, $2Ns$. When we doubled the population size to $N = 20{,}000$ and halved the selection coefficient s we observed essentially identical results (data not shown). The mutation coefficient we used was $\mu = 10^{-8}$ (rate of mutations per base pair, per meiosis). The recombination rate was 10^{-8} per base pair per meiosis. Figure 1 shows the simulation scheme. The output of the simulations was 100 common SNPs spread evenly over a 200-kb region (one SNP per 2 kb) and a variable number of causal SNPs (see Table 1) located within the central 9 kb of the 200-kb region.

The causal SNPs were subject to negative selection with selection coefficient s during the generation of the base population. Individuals carrying one or more of these variants have the disease with probability 0.1 (the penetrance). Individuals not carrying any of these variants have the disease with probability 0.01 (the sporadic rate). These parameters were chosen to obtain good power with reasonable sample sizes while being somewhat realistic. The ratio of penetrance to sporadic rate determines power. Thus increased power could be obtained either by increasing penetrance or decreasing the sporadic rate. If the penetrance is extremely high (such as 0.5) then strong familial aggregration will be seen and a family-based sampling approach would be more appropriate than a population-based sample. On the other hand, very low sporadic rates are unrealistic for complex diseases because the sporadic rate includes both non-genetic causes and genetic causes at other loci.

We consider a range of selection coefficients. If the selection coefficient is too low, most individuals will carry a causal variant, so the model considered here is not suitable. With $N = 10{,}000$ and $s = 0.0002$, a majority of individuals carry a causal variant. Thus, the smallest selection coefficient we consider is $s = 0.0005$, for which approximately 10–20% of individuals carry a causal variant, resulting in a prevalence

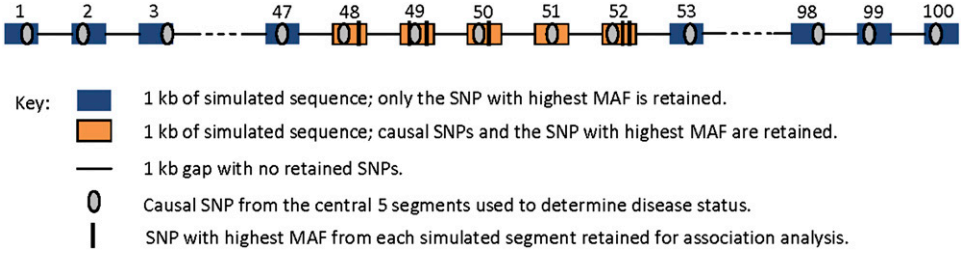

Figure 1 Simulation scheme. Each simulated region is made up of 100 simulated segments of length 1 kb with gaps of length 1 kb between them. The central five segments can contain causal SNPs. Causal SNPs are those that the simulation program designates as protein-changing mutations. These SNPs have been subject to negative selection at a specified rate. Only the causal SNPs and one SNP per segment with highest minor allele frequency (MAF) are retained. The causal SNPs are used to determine disease status, while the high MAF SNPs are tested in the association analysis. IBD status is determined through further simulation, as described in the main text.

of approximately 2–3% for the disease. If the selection coefficient is very high, the causal variants will tend to be very rare, and very large sample sizes are required to detect association, particularly with the SNP association test. The largest selection coefficient that we consider is 0.005. For each selection coefficient we generated 100 independent data sets, which were used as base populations for all the forward simulations described below. Table 1 shows the distributions of frequencies of causal variants and numbers of causal variants from the simulations.

After generating a base population with SFS_CODE, we ran our own forward simulator for 25 or 100 generations without selection or mutation or recombination, and with a different population size corresponding to a recently expanded population ($N_{\text{recent}} = 100{,}000$) or a population that has gone through a recent bottleneck ($N_{\text{recent}} = 1000$). Omission of selection, mutation, and recombination during this 25 or 100 generation time period is not significant because of the short timescale. Our forward simulator generated each new generation by sampling with replacement from the existing pool of haplotypes. When there were no causal variants segregating in the population after the final generation, the sampling process was rerun.

The purpose of the final generations of simulation using our own forward simulator was to determine IBD status. Haplotypes deriving from the same ancestral haplotype G generations ago (where $G = 25$ or $G = 100$) are considered to be detectably IBD (see Appendix). Current methods for IBD detection from SNP data can detect a majority of IBD segments arising from a shared ancestor within the past 25 generations. With improving SNP data and further developments in IBD detection methodology, it may soon be possible to detect IBD arising from shared ancestry 50 or even 100 generations ago.

In real data, ability to detect IBD depends on the number of generations to the common ancestor only through the length of the IBD segment. More recent common ancestry tends to result in longer IBD segments, which are easier to detect. However, the distribution of IBD lengths given the number of generations to the common ancestor is highly variable as it approximately follows an exponential distribution. Using a cut-off in terms of number of generations to common ancestor in place of a cut-off in terms of length of region simplifies the simulation procedure and gives some sense as to the properties of analysis of real data. When comparing a threshold of $G = 25$ generations in simulated data with a threshold of 2 cM in real data, say, on the one hand some IBD segments due to common ancestry 25 generations ago will be too short to be detectable in real data, while on the other hand some segments due to ancestry more than 25 generations ago will be long enough to be detected in real data, with these two effects cancelling each other to some extent.

Cases and controls were generated by sampling with replacement from the final generation until sufficient numbers were obtained. The numbers of cases and controls were chosen to achieve at least moderate power for both SNP association and IBD tests. In many instances the number of cases is much larger than would seem reasonable given the effective population size. However, in real life the actual population size is typically larger than the effective population size. Also, many human populations have undergone rapid expansion in the past few generations (due to improved medical care, etc), which again allows for realistically

Table 1 Properties of simulated causal variants

s	No. of variants	Variant frequencies	Haplotype carrier frequencies	Max R^2
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	0.000088–0.0011	0.0045–0.011	0.03–0.16

Interquartile ranges (IQR; lower quartile to upper quartile) from the 100 simulations with selection coefficient s are shown for several quantities of interest. The second column gives the number of causal variants per simulation. The third column gives the frequencies of the causal variants. The fourth column gives the proportion of haplotypes that carry a causal variant. The final column gives the maximum squared correlation coefficient between any one of the 100 common variants tested in the association test with any one of the causal variants. All results are from the base simulation population of 10,000 individuals.

increased sample sizes beyond the typical population size in the past 25–100 generations.

For the SNP association test, the retained common SNPs were tested for association with disease status using Fisher's exact allelic test. The minimum *P*-value over the 100 common SNPs was compared to a genome-wide significance threshold of 5×10^{-8}. The tested SNPs are common, while the causal SNPs tend to be rare. Table 1 shows that for low-selection coefficients many simulations contain at least one tested common variant that is very highly correlated with at least one causal variant, whereas for high-selection coefficients (very rare causal variants) the correlations tend to be low.

For the IBD test, the IBD status was determined by ancestry at the beginning of the final G generations of simulation, as described above. The difference between case/case and non-case/case IBD proportions was calculated, and 5 million permutations of case-control status were performed to obtain a *P*-value, which was compared to a genome-wide significance threshold of 2×10^{-6} for $G = 25$ or 5×10^{-7} for $G = 100$ (see the Appendix for derivation of these thresholds). Note that $G = 25$ corresponds approximately to the current resolution of IBD detection in SNP data, whereas $G = 100$ corresponds to improved resolution that may be achievable with denser data or improved methods. In the simulations the IBD status is the same at each of the simulated segments; thus a single IBD test covers the whole region.

For each parameter setting, 100 replicate data sets were generated using the 100 base populations generated with SFS_CODE. SNP association and IBD mapping tests were performed on the same data sets. The final generations of simulation were performed independently for the $G = 25$ and $G = 100$ generation simulations.

Table 2 shows results for a large recent effective population size ($N_{\text{recent}} = 100{,}000$). It can be seen that the IBD test with $G = 25$ is more powerful than the SNP association test when the selection against causal variants is sufficiently strong; however, once selection becomes too strong the variants are so rare that very large sample sizes are required to obtain reasonable power. The IBD test with $G = 100$ is at least as powerful as the SNP association test for all selection coefficients considered here. Within the model framework considered here, we cannot consider weaker selection and hence more common causal variants. SNP association testing would have superior power to IBD testing if one common causal variant dominated the effect of the gene on the disease.

By comparing results with $G = 25$ and $G = 100$ generations as the IBD detection cut-off we can see the effects of increased marker density on the power of IBD mapping. The $G = 25$ results correspond roughly to IBD detection with 500,000 common SNPs genome-wide. We do not yet have good data on the extent to which IBD detection power will improve with increased marker density and inclusion of lower frequency variants. The $G = 100$ results may correspond to very-high-density SNP data. The results in Table 2 show that improved IBD detection due to increased marker density has the potential to significantly expand the range of scenarios in which IBD mapping has more power than standard single-marker association testing. However, as noted in the introduction, IBD mapping cannot improve upon appropriate association testing (which may consist of rare-variant aggregation testing) in high-quality sequence data.

Table 3 shows results for a bottleneck that occurred 25 generations ago, when we can detect IBD arising from shared ancestry within the past $G = 25$ generations. The IBD test is more powerful than the SNP association test for all values of selection considered in this recent bottleneck scenario. Unless a causal variant is very common, it will tend to be represented by only a very small number of initial haplotypes and thus the IBD test has good power. On the other hand, during the time that the population is small, some haplotypes are lost and thus linkage disequilibrium increases between common SNPs and remaining haplotypes. This increases the power of the SNP association test, although not to a large enough extent to match the IBD test.

The situation in which the change in population size occurs at exactly the point in time at which the IBD reckoning begins is unrealistic. For the large population size, increasing the number of generations for which the population had the larger size to say 125 has little effect on either test because this is not enough time to add new

Table 2 Simulated power results: Large population size

s	No. of cases	No. of controls	Power assoc.	Power IBD25	Power IBD100	Assoc. *vs.* IBD25	Assoc. *vs.* IBD100
0.0005	500	500	0.87	0.57	0.81	assoc.	NS
0.001	500	500	0.65	0.53	0.81	NS	IBD
0.002	1000	1000	0.53	0.87	0.93	IBD	IBD
0.005	3000	3000	0.47	0.90	0.84	IBD	IBD

From an equilibrium population size of $N = 10{,}000$, the population was expanded to a recent effective size of $N_{\text{recent}} = 100{,}000$. The selection coefficient, s, used in simulating the equilibrium population is given in the first column. The second and third columns give the sample sizes. The fourth column gives the estimated power of the SNP association test with $G = 25$ generations at the recent effective population size; the power of the SNP association test with $G = 100$ generations was not significantly different (data not shown). The fifth and sixth columns give the estimated power of pairwise IBD tests with IBD determined from $G = 25$ and $G = 100$ generations at the recent effective population size, respectively. All power estimates are from 100 replicates; the standard errors are 0.03–0.05. The seventh column states whether the SNP association test or IBD test with $G = 25$ is more powerful if the difference is significant (two-sided paired *t*-test $P < 0.05$) or NS (nonsignificant) otherwise. Similarly the eighth column compares the SNP association test and IBD test with $G = 100$.

Table 3 Simulated power results: Small population size, very recent bottleneck

s	No. of cases	No. of controls	Power assoc.	Power IBD25	Assoc. *vs.* IBD25
0.0005	200	200	0.53	0.64	IBD
0.001	400	400	0.60	0.73	IBD
0.002	400	600	0.51	0.60	NS
0.005	400	1000	0.33	0.46	IBD

From an equilibrium population size of $N = 10{,}000$, the population was contracted 25 generations ago to a recent effective size of $N_{\text{recent}} = 1000$. The selection coefficient, s, used in simulating the equilibrium population is given in the first column. The second and third columns give the sample sizes. The fourth column gives the estimated power of the SNP association test, while the fifth column gives the estimated power of pairwise IBD test with IBD determined from the final $G = 25$ generations. All power estimates are from 100 replicates; the standard errors are 0.04–0.05. The sixth column states whether the SNP association test or IBD test is more powerful if the difference is significant (two-sided paired t-test $P < 0.05$) or NS (nonsignificant) otherwise.

haplotypes, and haplotypes do not tend to get lost in large populations over such relatively short timescales. On the other hand, for the small recent population size, increasing the amount of time over which the population is small has little effect on the power of the IBD test but significantly increases the power of the SNP association test because linkage disequilibrium is increased. Table 4 shows results with 100 additional generations (125 generations total) at size $N_{\text{recent}} = 1000$. Selection was applied during the forward simulation for these simulations. The power of the SNP association test is higher than that with the recent bottleneck, while the change in the power of the IBD test is not significant. As a result, the SNP association test is more powerful than the IBD test, in contrast to the recent bottleneck.

Analysis of Wellcome Trust Case Control Consortium (WTCCC) type 1 diabetes data

We analyzed the Wellcome Trust Case Control Consortium (WTCCC) type 1 diabetes data (Wellcome Trust Case Control Consortium 2007) to determine whether we could detect signals with IBD mapping that could not be detected with SNP association testing in these data. This data set is fairly large and homogeneous, with approximately 3000 controls and 2000 cases, all of whom are of European ancestry from the United Kingdom. Type 1 diabetes is a relatively good candidate for IBD mapping: it has relatively low prevalence (≈0.4%; Mehers and Gillespie 2008), and high heritability (≈88% heritability on the liability scale; Hyttinen *et al.* 2003). The major genetic contribution comes from the HLA (human leukocyte antigen) region, but other genes have also been implicated through association studies (Barrett *et al.* 2009). Multiple rare variants in one gene, *IFIH1* (interferon induced with helicase C domain 1), are associated with the disease (Nejentsev *et al.* 2009). However, the *IFIH1* variants are protective, which will tend to increase IBD in control/control pairs, while our IBD test looks for increased IBD in case/case pairs.

Before performing the analysis, we called the genotypes from the signal intensity data using BeagleCall, as described previously (Browning and Yu 2009). Highly accurate genotypes are critical for detecting IBD, and BeagleCall utilizes linkage disequilibrium information to significantly improve the genotype accuracy. Stringent quality control filters were also applied to the markers during the calling process. In total, 458,204 autosomal SNPs were analyzed in 1963 cases and 2938 controls.

The fastIBD module of Beagle 3.3.1 (Browning and Browning 2011) was used to detect IBD segments. We combined results from 10 runs of the program and used a threshold on the fastIBD score of 10^{-10}. In total, 4.1 million segments were found, or 0.42 per pair of individuals (0.1 per pair of haplotypes) over the autosome. The average number of SNPs covered by a segment was 464, while the average segment length was 3.2 cM (genetic distances obtained by interpolation from HapMap estimates; McVean *et al.* 2004). Figure 2 shows the distribution of lengths of detected segments. The genome average proportion of pairs IBD at a given location was 0.0003441 for non-case/case pairs and 0.0003447 for case/case pairs.

To perform the IBD test, we calculated the difference in IBD proportions between case/case pairs and non-case/case pairs and compared this difference to that obtained from 5 million permutations of case-control status. We calculated this difference and the corresponding permutation P-value at every tenth SNP along the autosomes. In addition, we

Table 4 Simulated power results: Small population size, older bottleneck

s	No. of cases	No. of controls	Power assoc.	Power IBD25	Assoc. v. IBD25
0.0005	200	200	0.71	0.55	Assoc.
0.001	400	400	0.76	0.67	NS
0.002	400	600	0.76	0.57	Assoc.
0.005	400	1000	0.73	0.51	Assoc.

From an equilibrium population size of $N = 10{,}000$, the population was contracted 125 generations ago to a recent effective size of $N_{\text{recent}} = 1000$. The selection coefficient, s, used in simulating the equilibrium population is given in the first column. The second and third columns give the sample sizes. The fourth column gives the estimated power of the SNP association test, while the fifth column gives the estimated power of pairwise IBD tests with IBD determined from the final $G = 25$ generations. All power estimates are from 100 replicates; the standard errors are 0.04–0.05. The sixth column states whether the SNP association or IBD test is more powerful if the difference is significant (two-sided paired t-test $P < 0.05$) or NS (nonsignificant) otherwise.

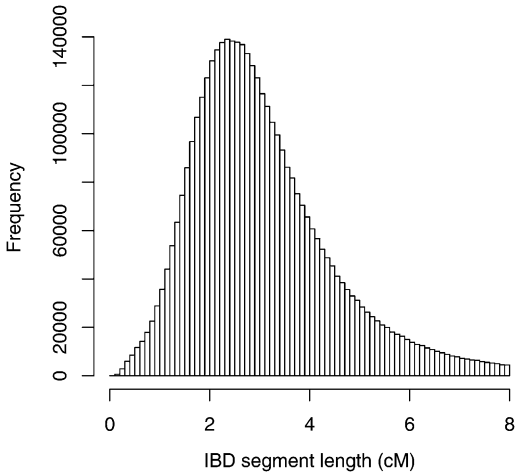

Figure 2 Distribution of lengths of detected IBD segments in the WTCCC type 1 diabetes data. IBD segments were detected using BEAGLE fastIBD. Lengths greater than 8 cM are not shown.

calculated permutation *P*-values genome-wide for 1000 permutations of case-control status, which allows us to determine the correct multiple-testing adjustment. The fifth percentile of the distribution of minimum *P*-value over the autosomes for the permuted data were 5.8×10^{-6}. This may be compared to the approximate multiple-testing adjusted thresholds calculated in the Appendix: 2×10^{-6} for $G = 25$ (corresponding to the approximate resolution of detection of 2 cM), which equates to a 2.2% genome-wide significance level in these data; or 4×10^{-6} for $G = 15$ (corresponding to the mean detected IBD segment length of 3.2 cM), which equates to a 3.4% genome-wide significance level in these data. Thus, use of the approximate genome-wide *P*-value thresholds calculated in the Appendix would be somewhat conservative in these data.

Figure 3 shows the unadjusted *P*-values. Because of the limited number of permutations, the smallest achievable *P*-value is $2 \times 10^{-7} = 1/(5 \times 10^6)$. The HLA region is clearly significant; however, this is not surprising given the extremely strong signal in this region. A region on chromosome 2 is almost significant (genome-wide adjusted *P*-value 0.20). A recent review (Baker and Steck 2011) lists two known associations with type 1 diabetes on chromosome 2. These are *IFIH1* at 2q24.2 and *CTLA4* (cytotoxic T lymphocyte associated antigen 4) at 2q33.2. The closer of these is *IFIH1*, which is 103 Mb away from IBD signal. The closest gene to the IBD signal is *BCL11A* (B-cell CCL/lymphoma 11A), which is 1.0 Mb away from the location of the smallest *P*-value on chromosome 2. *BCL11A* has been suggestively associated with type 2 diabetes (Zeggini *et al.* 2008) and affects pancreatic β-cell function (Simonis-Bik *et al.* 2010).

In the original SNP association analysis of these data (Wellcome Trust Case Control Consortium 2007), four loci were significant at a genome-wide significance threshold of 5×10^{-8}. These included the HLA region, with a *P*-value of 2×10^{-134}. In our IBD mapping analysis, the HLA region achieved the smallest *P*-value possible with the limited number of permutations performed ($P = 2 \times 10^{-7}$). Computational constraints preclude performing further permutations to determine how small a *P*-value can be obtained with IBD mapping in this region. Nevertheless, it is the achievement of genome-wide significance, rather than the actual *P*-value, that is important. Both IBD mapping and SNP association testing achieved genome-wide significance in the HLA region. However, SNP association testing was able to find genome-wide significant association at further loci, while IBD testing did not. Thus, SNP association testing found more significant results than the IBD testing in these data.

Discussion

IBD mapping has potential advantages over standard SNP association testing for detecting associations with rare variants using SNP array data, particularly when multiple rare variants are clustered within a gene. We showed through simulations that there are scenarios in which IBD mapping has higher power than SNP association testing. In large outbred populations, IBD mapping can have higher power when the level of selection against causal variants is relatively high, so that most causal variants remaining in the population are of recent origin. Because such variants are rare, either they must have strong effect sizes or there must be multiple causal variants clustered within a gene to have power to detect the association with a reasonably sized sample.

Our simulations also showed that IBD mapping has higher power than SNP association testing in founder populations provided that the founding event was very recent. If the number of generations since the founding event is approximately the same as the number of generations *G* for which shared ancestry can be detected, IBD mapping has higher power. On the other hand, if the founding event was further back, the SNP association test tends to have higher power because the additional generations of low population size significantly increase the level of linkage disequilibrium in the population.

The multiple testing correction for IBD mapping is lower than that for SNP association testing, and we derived genome-wide significance levels for IBD mapping which we used in our simulations and which could be used in analysis of real data if permutation-based adjustment is not feasible. The precise multiple testing adjustment depends on the sensitivity of IBD detection, which depends on the IBD detection method and on the characteristics of the data set, such as the SNP density. In the type I diabetes WTCCC data, the correct genome-wide significance threshold based on permutation was higher (less stringent) than the theoretical value that we had derived. Inference of IBD in real data are influenced primarily by the length of the IBD segment, while in our theoretical work IBD detection is based on depth of coancestry. A significant proportion of IBD segments due to common ancestry $G = 25$ generations ago will be shorter

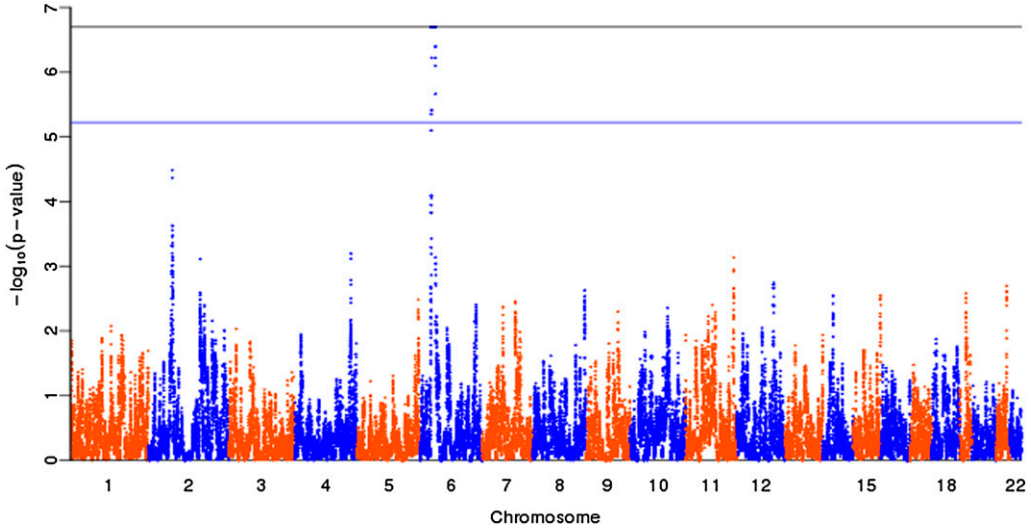

Figure 3 Permutation *P*-values for the IBD test in the WTCCC type 1 diabetes data. *P*-values were calculated at every tenth marker along the autosomes. The smallest possible *P*-value from the 5,000,000 permutations (2×10^{-7}) is shown by the black horizontal line. The genome-wide significance level determined by 1000 permutations (6×10^{-6}) is shown by the blue horizontal line.

than 2 cM, and in real data many such segments will not be detectable. The presence of very short IBD segments in the theoretical data leads to a faster rate of transitions in the IBD process, and hence to the need for a greater multiple testing correction, which may explain the difference between the theoretical and real data results.

In the WTCCC type 1 diabetes data we were able to detect an association with the HLA region using IBD mapping. However, given the huge effect of the HLA region, which is easily detected through SNP association tests, this was not an outstanding achievement.

Overall, while it is possible that IBD mapping will be useful in some circumstances, it seems doubtful that it will be worthwhile to routinely apply this approach. The best scenario for IBD mapping is a disease that has allelic heterogeneity (multiple causal variants within a gene, to provide an advantage relative to single-marker association testing), with low frequency causal variants (due to negative selection, so that shared causal variants are likely to be of recent origin and hence detectably IBD) and high heritability (high effect sizes, for reasonable power with moderate sample sizes). Our simulation results showed that IBD mapping can have higher power than association mapping in the case of moderately strong negative selection and allelic heterogeneity. Diseases with very high heritability are best suited to large-family studies, because ascertainment of large families increases allelic homogeneity (within families) and reduces the incidence of sporadic cases, thus increasing power (Wijsman and Amos 1997). However, when heritability is only moderately high, or when collection of family data are not practical, IBD mapping captures some of the advantages of the family-based approach.

It may be that the advantages of IBD mapping would be realized for very high sample sizes. IBD mapping is targeted at rare variants, and, particularly for very rare variants, large sample sizes are needed to see each such variant more than once in the sample. Rare variants that occur only once in a sample do not contribute to the IBD mapping statistic. Unfortunately, one needs to have a fairly homogeneous sample for IBD mapping, which is likely to preclude increasing sample sizes to high levels. The existence of population structure within a sample can cause false-positive IBD detection, and can also result in large differences genome-wide in IBD rates between case/case pairs and non-case/case pairs. Such structure need not be continental-level differences but can be, for example, Wales *vs.* England (Browning and Browning 2011). Although genome-wide differences can be subtracted out of the analysis, localized differences due to differences in informativeness of the haplotypes present in the populations may remain and cause false-positive IBD mapping signals.

As SNP array data become more dense, with arrays of several million SNPs, IBD segment detection will improve, and it will be possible to detect IBD due to more distant ancestry. This increases the range of scenarios over which IBD mapping can have good power, although such data will also have improved power for standard association testing. With high-quality sequence data, IBD mapping becomes irrelevant. This is because in the context of association testing (mapping), the usefulness of inferred IBD segments is to provide information about untyped variants in the vicinity of genotyped SNPs.

Acknowledgments

A full list of the investigators who contributed to the generation of the Wellcome Trust Case Control Consortium data are available from http://www.wtccc.org.uk. This study makes use of data generated by the Wellcome Trust Case Control Consortium and the Wellcome Trust Sanger Institute. Funding for the Wellcome Trust Case Control Consortium project was provided by the Wellcome Trust under award 076113. This work was supported by National Institutes of Health (NIH) awards R01HG005701, R01GM057091, and R37GM046255. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Wellcome Trust.

Literature Cited

Albrechtsen, A., T. S. Korneliussen, I. Moltle, T. V. Hansen, F. C. Nielsen *et al.*, 2009 Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet. Epidemiol. 33(3): 266–274.

Aldous, D. J., 2010 *Probability Approximations via the Poisson Clumping Heuristic*. Springer-Verlag, New York.

Baker, P. R., and A. K. Steck, 2011 The past, present, and future of genetic associations in type 1 diabetes. Curr. Diab. Rep. 11(5): 445–453.

Barrett, J. C., D. G. Clayton, P. Concannon, B. Akolkar, J. D. Cooper *et al.*, 2009 Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41(6): 703–707.

Bercovici, S., C. Meek, Y. Wexler, and D. Geiger, 2010 Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. Bioinformatics 26(12): i175–i182.

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 (in press).

Browning, B. L., and S. R. Browning, 2008 Haplotypic analysis of Wellcome Trust Case Control Consortium data. Hum. Genet. 123(3): 273–280.

Browning, B. L., and S. R. Browning, 2011 A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88(2): 173–182.

Browning, B. L., and Z. Yu, 2009 Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. 85(6): 847–861.

Browning, S. R., 2006 Multilocus association mapping using variable-length Markov chains. Am. J. Hum. Genet. 78(6): 903–913.

Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86(4): 526–539.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics 55(4): 997–1004.

Dudbridge, F., and A. Gusnanto, 2008 Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. 32(3): 227–234.

Feingold, E., 1993 Markov-processes for modeling and analyzing a new genetic-mapping method. J. Appl. Probab. 30(4): 766–779.

Francks, C., F. Tozzi, A. Farmer, J. B. Vincent, D. Rujescu *et al.*, 2010 Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. Mol. Psychiatry 15(3): 319–325.

Gao, X., L. C. Becker, D. M. Becker, J. D. Starmer, and M. A. Province, 2010 Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. 34(1): 100–105.

Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler *et al.*, 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19(2): 318–326.

Gusev, A., E. E. Kenny, J. K. Lowe, J. Salit, and R. Saxena *et al.*, 2011 DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. 88(6): 706–717.

Han, L., and M. Abney, 2011 Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. 35(6): 557–567.

Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24 (23): 2786–2787.

Houwen, R. H. J., S. Baharloo, K. Blankenship, P. Raeymaekers, J. Juyn *et al.*, 1994 Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat. Genet. 8(4): 380–386.

Hyttinen, V., J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto, 2003 Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes 52(4): 1052–1055.

Kong, A., G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich *et al.*, 2008 Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. 40(9): 1068–1075.

Leibon, G., D. N. Rockmore, and M. R. Pollak, 2008 A SNP streak model for the identification of genetic regions identical-by-descent. Stat. Appl. Genet. Mol. Biol. 7(1): 16.

Li, B., and S. M. Leal, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83: 311–321.

Lowe, J. K., J. B. Maller, I. Pe'er, B. M. Neale, J. Salit *et al.*, 2009 Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. PLoS Genet. 5(2): e1000365.

Madsen, B. E., and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5(2): e1000384.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461(7265): 747–753.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science 304(5670): 581–584.

Mehers, K. L., and K. M. Gillespie, 2008 The genetic basis for type 1 diabetes. Br. Med. Bull. 88(1): 115–129.

Moltke, I., A. Albrechtsen, T. V. Hansen, F. C. Nielsen, and R. Nielsen, 2011 A method for detecting IBD regions simultaneously in multiple individuals—with applications to disease genetics. Genome Res. 21(7): 1168–1180.

Nejentsev, S., N. Walker, D. Riches, M. Egholm, and J. A. Todd, 2009 Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925): 387–389.

Pe'er, I., R. Yelensky, D. Altshuler, and M. J. Daly, 2008 Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. 32(4): 381–385.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3): 559–575.

Simonis-Bik, A. M., G. Nijpels, T. W. van Haeften, J. J. Houwing-Duistermaat, and D. I. Boomsma *et al.*, 2010 Gene variants in the novel type 2 diabetes loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B affect different aspects of pancreatic beta-cell function. Diabetes 59(1): 293–301.

Te Meerman, G. J., M. A. Van Der Meulen, and L. A. Sandkuijl, 1995 Perspectives of identity by descent (IBD) mapping in founder populations. Clin. Exp. Allergy 25: 97–102.

Thomas, A., N. J. Camp, J. M. Farnham, K. Allen-Brady, and L. A. Cannon-Albright, 2008 Shared genomic segment analysis: mapping disease predisposition genes in extended pedigrees using SNP genotype assays. Ann. Hum. Genet. 72: 279–287.

Thompson, E. A., 2009 Inferring coancestry of genome segments in populations. Invited Proceedings of the 57th Session of the International Statistical Institute, IPM13. Durban, South Africa.

Van der Meulen, M. A., and G. J. te Meerman, 1997 Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet. Epidemiol. 14(6): 915–919.

Voight, B. F., and J. K. Pritchard, 2005 Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1(3): e32.
Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661–678.
Wijsman, E. M., and C. I. Amos, 1997 Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet. Epidemiol. 14(6): 719–735.
Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.
Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89(1): 82–93.
Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini *et al.*, 2008 Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. 40(5): 638–645.
Zhu, X., T. Feng, Y. Li, Q. Lu, and R. C. Elston, 2010 Detecting rare variants for complex traits using family and unrelated data. Genet. Epidemiol. 34(2): 171–187.

Communicating editor: H. Zhao

Appendix

In the theoretical work that follows we assume that the genome-wide case/case and non-case/case rates of IBD are the same. This assumption is wrong (Devlin and Roeder 1999; Voight and Pritchard 2005), but for a complex disease will not be too far off, provided the samples really do come from the same population. In analysis of real data we subtract the genomic averages.

Background rate of IBD

The expected rate, r_0, of IBD in controls depends on both the population (in particular its effective size) and the power to detect IBD. Power to detect IBD depends on the type of genetic data, such as dense SNP array or sequence data, and also on the algorithm used to detect the IBD. With the Beagle fastIBD algorithm (Browning and Browning 2011) with Affymetrix 500K array data in the UK population (WTCCC data), we found IBD at rate 3.4×10^{-4} per pair of individuals (or 8.5×10^{-5} per pair of haplotypes) in controls. Moreover we had high power to detect IBD segments of size 2 cM or larger while controlling the false discovery rate to be close to zero. With denser data we would expect to be able to find much of the IBD of size 1 cM or larger. IBD segments of size 1 cM are the expected size for common ancestry 50 generations ago (100 meioses linking the pair of individuals) while segments of size 2 cM correspond to 25 generations.

We note that the length of an IBD segment deriving from a common ancestor G generations ago has approximately an exponential distribution and thus has high variance. In practice, IBD detection depends on the length of the segment, rather than on the number of generations to the common ancestor. However, considering IBD detectability in terms of the number of generations greatly facilitates the theoretical analysis.

Let N be the effective size of the population (the equivalent number of randomly mating individuals, which is generally smaller than the actual population size). Let f_g be the probability that two randomly chosen alleles at generation 0 (now) are IBD if all alleles are non-IBD at generation g (g generations into the past). Then (Wright 1931)

$$f_g = 1 - \left(1 - \frac{1}{2N}\right)^g \approx 1 - e^{-g/(2N)} \approx g/(2N).$$

Suppose we can detect IBD due to common ancestry back to G generations ago. Then the approximate detectable rate of IBD between haplotypes will be $1 - \exp(-G/(2N)) \approx G/(2N)$. For example, if $G = 25$ and $N = 1.5 \times 10^5$, the approximate IBD rate for pairs of haplotypes is 8.3×10^{-5}, which is similar to the level we found in the United Kingdom (Browning and Browning 2011).

Genome-wide significance level for IBD testing

In this section we derive an approximate genome-wide significance level for IBD mapping. First consider only the IBD process in the case/case samples. That is, assume that we know the background rate of IBD and want to know whether the case/case rate of IBD is significantly higher at some point in the genome. We also assume that the rate of IBD along the chromosome is constant under the null hypothesis of no causal variants. To match notation used elsewhere, we refer to distance along the chromosome as "time." Thus, in this section "time" refers to genetic distance (measured in centimorgans) rather than to number of generations. Following Aldous (2010) (chapter B, Markov chain hitting times) and Feingold (1993) we approximate the distribution of the hitting time of the IBD process with an exponential distribution.

Let X_t be the number of case/case haplotype pairs that are IBD at position t (position measured in genetic distance) on the chromosome of interest. As an approximation, assume that X_t is a Markov process with instantaneous transition rates $Q(i, i + 1) = \lambda_0(n - i)$ and $Q(i, i - 1) = \lambda_1 i$, where n is the number of case/case haplotype pairs sampled. That is, each non-IBD pair becomes IBD at rate λ_0 (moving along

the chromosome) while each IBD pair becomes non-IBD at rate λ_1. Let T_b be the location of the first point of the chromosome at which X_t reaches or exceeds a threshold b, where b is sufficiently large so that $P(X_t \geq b)$ is small. Around the point b, we approximate the process by an asymmetric random walk, with $Q(i, i+1) = \lambda_0(n-b)$ and $Q(i, i-1) = \lambda_1 b$. For such a random walk, the expected sojourn time (distance) in state b is $S(b) = (\lambda_1 b - \lambda_0(n-b))^{-1}$. This sojourn time represents the aggregation of several visits over a short time period (distance), or a "clump." The expected time (distance along the chromosome) between such clumps is also the expected hitting time (assuming an exponential distribution) and is thus $E(T_b) = S(b)/\pi(b)$, where $\pi(b)$ is the proportion of time spent in state b. For the Markov process described above, $\pi(b)$ is binomial $(n, \lambda_0/(\lambda_0 + \lambda_1))$. Note that this makes it clear that we are assuming that the IBD status of pairs are independent, which is not fully correct, but is an approximation for small expected IBD proportion $r_0 = \lambda_0/(\lambda_0 + \lambda_1)$. Then the distribution of T_b is approximately exponential with rate

$$1/E(T_b) = (\lambda_1 b - \lambda_0(n-b))\binom{n}{b}(1-r_0)^{n-b}r_0^b$$

Thus, given a total genetic length L, the multiple-testing adjusted P-value corresponding to an observed number of IBD pairs b is

$$\begin{aligned} P_{\text{adj}} &= P\left(\max_{0\leq t\leq L} X_t \geq b\right) \\ &= P(T_b \leq L) \\ &= 1 - \exp\left\{-L(\lambda_1 b - \lambda_0(n-b))\binom{n}{b}(1-r_0)^{n-b}r_0^b\right\}. \end{aligned}$$

The unadjusted P-value is a binomial probability

$$P_{\text{unadj}} = P(X_t \geq b),$$

where X_t follows a Binomial(n,r_0) distribution. If we find a value of b that gives an adjusted P-value of approximately 0.05, the corresponding unadjusted P-value (for the same value of b) gives a genome-wide significance threshold for unadjusted P-values.

For a value of G that gives the resolution of IBD detection (IBD with a common ancestor within the past G generations), and with an effective population size of N_e, the rate of IBD detection is $r_0 = G/(2N_e)$ (derived in the previous section). We could assume that the average size of IBD tract is $100/(2G)$ cM, giving $\lambda_1^* = G/50/\text{cM}$. This fits the exponential assumption, but ignores the fact that most IBD tracts actually result from a common ancestor occurring within $<G$ generations. In fact, given that a pair of haplotypes are IBD at a given position (with common ancestor within the past G generations), the time to the most recent shared ancestor is approximately uniform on $1 \ldots G$ (assuming constant population size). The length of IBD tract is then not exponential, but the average rate out of IBD is $\lambda_1^{**} = \sum_{t=1}^{G}(1/G)t/50 = (G+1)/100/\text{cM}$. The rate λ_0 from non-IBD to IBD can be found by solving $r_0 = \lambda_0/(\lambda_0 + \lambda_1)$. Hence $\lambda_0 = \lambda_1 r_0/(1 - r_0)$.

For each value of G considered, we investigated the value of P_{unadj} that gives a value of $P_{\text{adj}} = 0.05$. We tried various values of effective population size N and large values of number of case/case haplotype pairs n. We saw essentially no effect of effective population size, but slightly higher (less stringent) thresholds for smaller sample sizes. The values given are for very large sample sizes (*e.g.*, 400 million pairs of haplotypes, corresponding to approximately 14 thousand individuals). We used $L = 3000$ cM. For λ_1^* we get a genome-wide P-value threshold of approximately 9.1×10^{-6} for $G = 5$, 2.7×10^{-6} for $G = 15$, 1.5×10^{-6} for $G = 25$, 7.1×10^{-7} for $G = 50$ and 3.3×10^{-7} for $G = 100$. For λ_1^{**} the thresholds are slightly higher (less stringent) because the IBD segments tend to be slightly longer: 1.6×10^{-5} for $G = 5$, 5.3×10^{-6} for $G = 15$, 3.1×10^{-6} for $G = 25$, 1.5×10^{-6} for $G = 50$, and 6.9×10^{-7} for $G = 50$.

We now return to the question of the IBD test statistic, which compares case and control rates of IBD. If X_1 is the number of IBD case/case haplotype pairs and X_0 is the number of IBD non-case/case haplotype pairs, then each has approximately the distribution described above. Moreover the behavior of the normalized values (X_i divided by the number of pairs of haplotypes interrogated) will have the same properties in terms of P-value adjustment. Since the P-value adjustment does not depend significantly on sample size, and the case/case pairs are approximately independent of the non-case/case pairs, the difference of the normalized values should also have similar properties in terms of P-value adjustment.

It is important to emphasize that the discussion in this section involves a great deal of approximation. Nonetheless, the results look reasonable and will be useful for the purposes of comparing power. On the basis of the results presented above, we suggest a genome-wide P-value threshold of 1×10^{-5} for $G = 5$, 4×10^{-6} for $G = 15$, 2×10^{-6} for $G = 25$, 10^{-6} for $G = 50$, and 5×10^{-7} for $G = 100$.