European Journal of Human Genetics (2007) 15, 1260–1268

www.nature.com/ejhg

ARTICLE

A simulation-based analysis of chromosome segment sharing among a group of arbitrarily related individuals

Ondrej Libiger[1,2,3] and Nicholas J Schork*[,1,2,3]

[1]*Scripps Genomic Medicine, Scripps Health, La Jolla, CA, USA;* [2]*Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, USA;* [3]*Center for Human Genetics and Genomics, University of California at San Diego, La Jolla, CA, USA*

A fundamental set of issues in human genetics research concerns the statistical properties of the DNA sequence or chromosomal segments that are shared between related individuals. Although well-established mathematical formulations exist that consider such sharing via measures such as the kinship coefficient, many of these formulations are derived for entire genomes, individual sequence variations, or small stretches of DNA, and hence, do not consider either the actual size or the number of the genome-wide chromosomal segments that are shared between two or more arbitrarily related individuals. In this paper, we employ a flexible gene-dropping simulation-based approach for estimating the distribution of the size and the number of chromosomal segments shared by any number of arbitrarily related individuals. The approach takes advantage of chromosome- and sex-specific recombination rates adopted from integrated genetic and physical maps, and considers the genome as a whole, rather than specific genomic regions or loci. In addition, our analysis considers the effects of linkage disequilibrium and crossover interference on segment sharing. Our proposed analysis and computational strategy can be used to provide compelling answers to questions concerning variation in the kinship coefficient as well as the distribution of chromosomal sharing over individual chromosomes. We present results that showcase possible application of assessing genomic sharing in gene mapping and apply our analysis to data available from published gene mapping studies.
European Journal of Human Genetics (2007) **15,** 1260–1268; doi:10.1038/sj.ejhg.5201910; published online 15 August 2007

Keywords: chromosome segment sharing; identity-by-descent; kinship; linkage disequilibrium; crossover interference

Introduction

Probabilistic analysis of the fraction of genome shared by individuals known to have genealogical links has a central place in genetic research. For example, gene mapping strategies such as pedigree-based linkage analysis and haplotype-based linkage disequilibrium (LD) analysis often require and/or exploit knowledge of the probability that a pair of individuals share chromosomal segments of a certain size. One fundamental concept in genetic research of relevance to genome sharing is the 'kinship coefficient', or the probability that a gene taken at random from individual *i* at a given locus is identical-by-descent (IBD) to a gene taken at random from individual *j* at the same locus.[1] The kinship coefficient (or rather twice the kinship coefficient given that humans are diploid) at a single locus can be extrapolated to the genome as a whole, which

*Correspondence: Dr NJ Schork, Molecular and Experimental Medicine, Scripps Research Institute, MEM 275, 10550 Torrey Pines Road, La Jolla, CA 92037, USA. Tel: +858 784 2308; Fax: +858 784 2910;
E-mail: nschork@scripps.edu
Received 6 February 2007; revised 9 July 2007; accepted 20 July 2007; published online 15 August 2007

results in the expected fraction of the genome that is shared IBD between two individuals of known ancestry (eg, twice the kinship coefficient for siblings is 0.5 and thus siblings share, in expectation, one-half of their diploid genomes). Although theoretical formulae exist for calculating the kinship coefficient for pairs of relatives,[2] assessing the variance of the kinship coefficient has not received much attention.[3] In addition, the kinship coefficient considers only individual loci or expectation over the entire genome and therefore provides no information about the distribution of shared chromosomal segments in terms of both their size and number.

The need for statistical constructs that go beyond the single locus-oriented kinship coefficient for genetic analysis is most clearly apparent in gene mapping studies based on haplotype sharing. Haplotype-sharing analysis is often based on a probabilistic comparison of chromosome segment sharing among a group of related individuals possessing a certain trait or disease relative to the expected sharing among individuals without the trait or disease (for a detailed, albeit somewhat dated, description of haplotype-sharing analysis, see Schork *et al*,[4] and, for an updated list of references describing probabilistic studies of segment sharing, see Table 1). A number of researchers have developed statistical analysis methods for haplotype-based gene mapping studies. Table 1 and the discussion provided in Schork *et al*[4] summarize the literature investigating chromosomal segment sharing and/or the short-term 'evolution' of chromosomal segments in genealogically-defined populations (ie, a group of individuals whose genealogical links are known). The cited studies approach relevant issues from a wide variety of angles: some are based on theoretical calculations, while others use a 'gene-dropping' simulation approach. More recently, Chapman and Thompson[15] employed Monte Carlo Markov Chain models to investigate the number of IBD-shared segments between related individuals[14] as well as their size in isolated populations that originated from a recent yet small number of founders. They also considered the effect that population growth and subdivision have on this chromosomal segment sharing. Similar phenomena have been studied by Stefanov,[11] who investigated the cumulative probabilities of the proportion of shared genomic segments that are IBD, and, in follow-up work, considered the proportion of the genome, as well as the number of genomic segments containing IBD-shared haplotypes, via Monte Carlo Markov Chain models.[12] However, Stefanov's numerical evaluations are applicable only to full- and half-sibling pairs, grandparent–grandchild pairs, and great-grandparent/great-grandchild pairs.

There are additional issues that are important to consider in the evaluation of chromosomal segment sharing. For example, there is a need to recognize and evaluate the distinction between sharing of alleles at adjacent loci IBD against the sharing of alleles merely identical-by-state (IBS). To make compelling claims about the probability that individuals actually share a segment of chromosome IBD based on observed genotype data, one must compute the probability that the observed sharing of homozygous alleles or a single chromosomal segment actually reflects sharing of a common ancestral segment, and not merely the probability that the observed allele sharing associated with a string of adjacent multilocus genotypes occurs purely by chance. In computing the conditional probability that a set of individuals share a chromosomal segment IBD given that they share a string of adjacent genotypes or alleles IBS, a number of factors must be accommodated, such as, for example, the pedigree structure linking the individuals in question, the density of genetic markers providing genotype or allele data, the informativity (ie, allele frequencies) of the markers used, and recombination rates in the genomic region(s) harboring the marker loci.

Nolte and te Meerman[13] have recently described a simulation method for determining the probability that a segment shared by individuals within a population is IBD

Table 1 Selected theoretical and empirical studies published after 1998 investigating the short-term evolution of chromosomes (or chromosomal segments)

Authors	*Year*	*Pop*	*ST*	*Comments*
Genin *et al*[5]	1998	P	A	Expected size of an autozygous segment around a disease locus
Schaffer[6]	1999	P	A	Algorithm for computing interval probability of autozygosity
Broman and Weber[7]	1999	O	H	Empirical size of homozygous segments
Clark[8]	1999	O	H	Model for homozygous segment size distribution
Wiuf and Hein[9]	1999	O	IBD	Ancestry of sequences sampled from the coalescent with recombination
McPeek and Strahs[10]	1999	O	IBD	Assessment of linkage disequilibrium by the decay of haplotype sharing
Stefanov[11]	2000	SP	IBD	Cumulative probabilities of the proportion of shared genomic segments
Stefanov[12]	2002	SP	IBD	Proportion of shared genome and number of shared genomic segments
Nolte and te Meerman[13]	2002	O, I	IBD, IBS	Probability that a shared segment is identical by descent
Chapman and Thompson[14]	2002	O	IBD	Number of shared segments
Chapman and Thompson[15]	2003	I	IBD	Size of shared segments
Leal *et al*[16]	2005	P	IBS	Software computing probabilities of sharing identical by state

Abbreviations: Pop, population type (I, isolated population; O, outbred population; P, pedigrees; SP, simple pedigrees); ST, segment type (A, diploid IBD homozygous segments; H, diploid IBS homozygous segments; IBD, identical by descent; IBS, identical by state).

given the population history. The genetic analysis software package MORGAN[17–19] can be used to assess the probability that a single locus or a fixed number of consecutive loci are IBD in the absence of marker data, as well as conditionally on observed marker data. However, usage of MORGAN to assess the number and size distribution of shared chromosomal segments in the entire genome would be very difficult and impractical since it was not necessarily designed for this purpose. MORGAN is also limited in the number of individuals whose genomes can be analyzed for sharing.

A second set of issues in the analysis of chromosomal segment sharing concerns consideration of the effects of LD and crossover interference.[20–22] Accommodating the influence of crossover interference is particularly difficult since practical mathematical models for crossover interference that can be incorporated into relevant segment-sharing calculations and/or simulation studies do not exist.

To date, no study has been published that considered genomic sharing within a group of arbitrarily related individuals within the genome as a whole, and therefore could provide a comprehensive view of chromosomal segment sharing in pedigrees. Furthermore, no tools are available that can easily perform such analyses. Our study is aimed at filling this gap. We have employed a flexible genome-wide gene-dropping simulation-based approach to assess chromosomal segment sharing throughout the genome for an arbitrary number of arbitrarily related individuals. This procedure provides reliable estimates of expected IBD and IBS chromosome segment and genome-wide allele-sharing probabilities. Our approach enables us to determine whether a segment shared IBS is also shared IBD, since it maintains marker genotype data together with information about the parental origin of relevant alleles. A novel characteristic of our simulation method is that it takes advantage of integrated genetic and physical maps that contain chromosome-specific, empirically determined male and female recombination rates. This feature allowed us to produce not only genome-wide results, but also results tailored to a specific position in the genome. For example, two different chromosomal segments of equal physical size (ie, number of bases) may exhibit different sharing probabilities due to variable recombination rates throughout the genome. Such variability has been reported by several studies.[23–25] While many previous studies circumvent this issue by reporting the size of chromosomal segments in units of genetic distance, we feel that as sequence data become more readily available and the main focus shifts from markers to sequence, it is more advantageous to present results that concern chromosomal segments in units of physical distance, that is, in the number of bases. This approach also allowed us to drop objectionable assumptions concerning the use of mapping functions.[26–29] In addition, we have devised a scheme for use with our simulation approach that allows consideration of the effects of LD and crossover interference. We present applications of our methodology in a wide variety of settings.

Methods

The simulation procedure

Our approach to computing probabilities associated with chromosomal segment sharing is based on simulating recombination events and the transmission of gametes to offspring. Basically, we simulate the inheritance of a genome (ie, pairs of chromosomes in a diploid species) within a given genealogy. In this respect, our methodology is rooted in classical gene-dropping approaches to genetic simulation,[30] and is, in essence, similar to the approach that Nolte and te Meerman[13] employed for population analyses, and to the approaches that are implemented in the 'ibddrop' and 'autozyg' programs of the MORGAN package.[17–19] However, there are several unique characteristics of our strategy. First, it can be used to analyze sharing in entire genomes easily. Also, it can be used to assess the size distribution and number of all shared chromosomal segments in a given genomic region. Finally, our strategy can be used to assess sharing in a number of related individuals effectively. The time requirements of the algorithm grow linearly with the pedigree size and the number of markers. A sharing analysis in a nuclear family involving 30 000 biallelic markers and 100 simulation runs required approximately 30 min to complete on a PowerBook G4 laptop computer.

Our simulation methodology begins by assigning an array representing the genome to each founding member of the genealogy (ie, those individuals without parental information). This array contains two pieces of information: (1) a unique genome identifier at each genomic position for each of an individual's two chromosomes; and (2) the variants or alleles an individual possesses at each relevant location. At each of these locations, an allele is assigned to each individual that is consistent with either the actual marker allele that the individual possesses (eg, as taken from an observation made on that individual) or based on the allele frequencies associated with the population that the individual is assumed to have come from. The chromosome assignment process is different for founders (individuals without parents specified in the genealogy) and non-founders (those with specified parents); that is, founder chromosomes are assigned a unique number and marker alleles based on what is known about the alleles that individual possesses or is assigned via a random number generator, based on allele frequency information in the population. For non-founders, the simulation proceeds by first identifying all the genealogical links from parents to their offspring, beginning with the founders and continuing to the youngest generation. For each parental pair, the program simulates meiosis and

recombination by assigning alleles to their offspring (ie, 'non-founders') arrays based on what was transmitted to them after simulating meiotic events in their parents. This is accomplished by traversing the parents' arrays and calculating the probability of recombination occurring between adjacent loci. As described previously, we used recombination probabilities based on observed, empirically established, recombination fraction information associated with a specified genetic map. Once allele codes have been assigned to every member of the pedigree, the assigned chromosomes can be analyzed for sharing of alleles across individuals within the genealogy. The observed genotype or haplotype information of these individuals can be used to determine the extent and type of sharing to which sharing probabilities can then be assigned through the simulation. Our program was not designed to compute specifically the probability that arbitrary individuals in a pedigree share chromosomal segments conditional on particular non-founders having very specific multilocus genotypes, although the program could be used with rejection sampling to compute such probabilities.

Our simulation procedure can be used to differentiate genetic variants that are shared among individuals' chromosomes either IBS or IBD. Sharing of a marker allele at a polymorphic site (ie, specified marker locus) on the chromosomes of several individuals reflects (initial) IBS sharing. By determining whether the underlying chromosome identifier number associated with the polymorphic sites assessed for allele sharing is the same or not, one can examine if a chromosomal segment or a haplotype was actually inherited from a common ancestor (ie, whether or not the segment is shared IBD). Using this information, the method can be used to provide the conditional probability that a given segment is shared IBD when it is observed as IBS. For situations involving homozygosity mapping analyses, the computed probabilities relate to an observed homozygous segment and depend on whether that shared homozygous segment is autozygous or not. Required for the method is the genealogical information, information about relevant polymorphic loci and a marker map (ie, interlocus distances in base pairs and recombination rates), and marker allele frequencies.

The genetic and physical map

Empirical recombination rates were obtained from LDB2000,[31,32] an integrated genetic and physical map. This high-resolution map provides locations for 32 673 genes, polymorphisms, and sequence-tagged sites throughout the genome. The loci in the map were ordered by their approximate midpoint location in the consensus DNA sequence relative to an origin at the p-telomere. Linkage maps were based on pairwise logarithm of odds (LOD) score criteria computed with genotype data from the CEPH panel, version 9,[33] as well as lod scores associated with Genatlas.[34]

Interference model

We designed a model of crossover interference in which an occurrence of a recombination event between two genetic loci prohibits the occurrence of another recombination event between the locus that directly antecedes the original recombination event and a locus located less than r bases away, where r is the specified 'range of interference'. Thus, under this model, all recombination events occur at least $r-d$ bases apart, where d is the distance separating the two loci that flank the first recombination event. Since we used a very dense map (the mean interlocus distance on chromosome 1 is 74 kb), and set r to be relatively large (1–100 Mb), the interlocus distances were negligibly small compared to the parameter specifying the range of interference. Because our method simulates recombination events in one direction from the beginning of the chromosome toward its end, and recombination events occurring near the previous recombination event are prohibited, some chromosomal regions exhibit less recombination than the input recombination rates specify. We, therefore, generated a new set of recombination rates by increasing the original recombination rates extracted from the integrated map, and utilized these corrected recombination rates in the simulations whenever we employed our crossover interference model. Using the new interference-corrected recombination rates in our simulation program with the crossover interference model produced recombination counts between every pair of loci along the genome that corresponded to the original recombination rates.

Results

We showcase the utility of assessing relevant probabilities in a number of settings. First, we describe the assessment of variation in human kinship (ie, we explored the standard error of the kinship coefficient and its generalization to the entire genome), and the distribution of the size and number of shared chromosome segments. We present results using two example pedigrees consisting of a sibling and a cousin pair. We also describe applications of our method to analyses that involve a large number of arbitrarily related individuals in extended pedigrees, and we compute relevant probabilities associated with example gene mapping studies based on haplotype sharing. In these analyses, we also explore the effect of LD on these probabilities. We also present an application of chromosomal segment-sharing analysis to gene mapping studies that assess chromosomal segment-sharing information to assist in the design of such a study with respect to marker density. Finally, we explore the effects of our crossover interference model on shared chromosomal segment size and number as well as the overall fraction of genome that is shared between related individuals.

Variation in the kinship coefficient

We calculated the average fraction of shared genome that is IBD for different pairs of relatives (ie, a generalization of the kinship coefficient for pairs of individuals) as well as the SD of shared genome using our strategy. Some results of these simulations are provided in Table 2 and Figure 1, which describe the distribution of the fraction of genome shared by sibling pairs over the genome as a whole, and for chromosome 1 only, based on 100 000 simulation runs. Note that the center of the distribution for chromosome and genome sharing is 0.5 – twice the kinship coefficient of sibling pairs – as expected, but there is considerable variation in the fraction of shared genome. Thus, on average, sibling pairs will share 50% of their genomes; however, only ~75% of these pairs will actually share between 45 and 55% of their genomes IBD. In the context of single chromosomes, this variation is much larger, as can be seen in Figure 1. This reflects the fact that high-sharing fractions associated with some chromosomes in the genome will tend to negate the effect of low-sharing fractions associated with other chromosomes due to variable recombination rates and the different lengths of individual chromosomes and chromosomal regions.

Distribution of shared chromosomal segment sizes and numbers

We evaluated the distribution of chromosome segment sharing, both in terms of the average or expected number of shared segments and the average size of these shared segments, for different relative pairs. Table 2 provides the results of 1000 simulation runs used to evaluate the average number and size of the chromosomal segments shared IBD by sibling pairs. Our results for overall sharing (ie, sharing between two paternal chromosomes, two maternal chromosomes and sharing between paternal and maternal chromosomes) between a pair of siblings show that the expected proportion of chromosomes, as well as whole autosomal genome, that is shared is 50% as expected from theoretical calculations (eg, using twice the kinship coefficient).

We also considered the length of genomic regions harboring alleles shared IBS by these pairs. We did this by making some restrictive assumptions to showcase the method; we first assigned alleles at loci based on the locations described in the LDB2000 map,[31,32] and also assumed that each locus in this map had five equally frequent alleles. Further, we assumed that the marker loci were in linkage equilibrium. The results (Table 2) suggest that sharing of alleles IBS at a number of adjacent loci with these assumed allele frequencies has a high probability except when the number of adjacent loci is large.

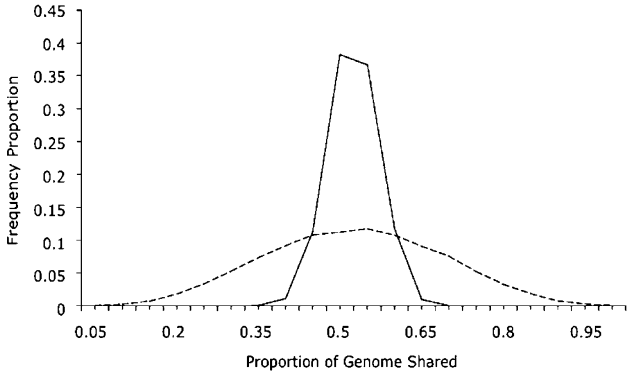

Figure 1 The distribution of IBD sharing between siblings. The distribution (ie, frequency histogram) of the amount of genome that is IBD in a pair of siblings based on 100 000 simulations. The solid line corresponds to the proportion of the whole autosomal genome, whereas the dotted line depicts the proportion of chromosome 1. Chromosome 1 was chosen arbitrarily to show increased variance in sharing in single chromosomes compared to the entire genome.

Table 2 Estimated example of chromosome and whole autosomal genome sharing parameters for sibling pairs

Chromosome	*Sharing*	*IBD (%)*	*IBS (%)*	*No. of IBD Segs*	*No. of IBS Segs*	*Ave IBD*	*Ave IBS*
1	Maternal	48.7±19.9	58.9±15.9	3.9±1.4	275.9±102.2	30.97±34.31	0.52±5.46
	Paternal	51.3±26.9	61.1±21.6	2.3±1	261.8±141.5	53.87±57.06	0.57±7.4
	Overall	50±16.9	68±10.9	6.2±1.7	771.1±253.2	39.61±45.65	0.43±5.42
6	Maternal	50.9±22.4	60.7±18	3±1.1	144.5±68.7	29.45±30.98	0.72±6.08
	Paternal	48.4±32	58.7±25.7	1.9±0.9	150.9±96.1	44.37±50.31	0.66±7.43
	Overall	49.7±19.1	67.8±12.3	4.8±1.4	423.4±164	35.22±40.25	0.55±5.7
21	Maternal	47±32.7	58.2±26.4	1.2±0.8	36.8±4.04	17.75±13.14	0.74±4.04
	Paternal	49.1±39	59.4±31.5	0.9±0.6	35.4±4.77	24.58±16.36	0.79±4.77
	Overall	48±25.9	67±16.7	2.1±1	103±3.74	20.69±15	0.61±3.74
Genome	Maternal	49.8±5.5	59.9±4.4	52±4.8	2658.6±291.9	27.46±28.62	0.65±5.54
	Paternal	49.8±7.2	59.9±5.8	36.4±3.8	2641.8±373.9	39.27±44.32	0.65±6.9
	Overall	49.8±4.4	67.9±2.8	88.4±6.1	7588.4±663.4	32.32±36.36	0.51±7.43

Abbreviations: Ave IBD, average size (in millions of bases) of shared chromosomal segments identical by descent (SD); Ave IBS, average size (in millions of bases) of shared chromosomal segments identical by state (SD); IBD, estimated percentage of chromosome or genome shared identical by descent (SD) : the variability of the means is caused by sampling variability; IBS, estimated percentage of chromosome or genome shared identical by state given allele frequency assumptions (SD); Segs, segments.

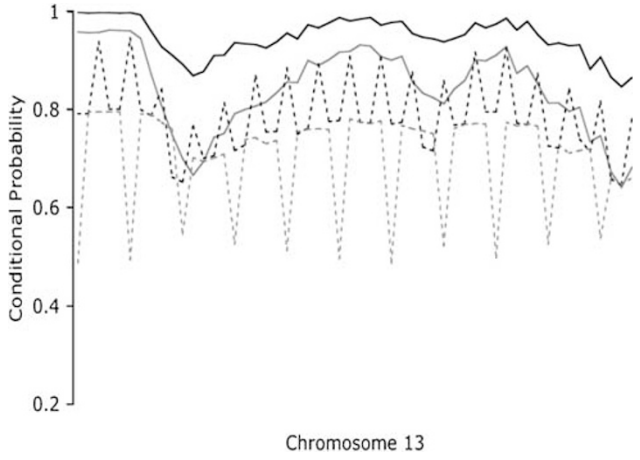

Figure 2 The probability of sharing a 10 Mb IBD segment when IBS sharing is observed with markers 2 Mb apart (chromosome 13). The conditional probability that affected members of the Mennonite pedigree (depicted in Figure 1 in Puffenberger *et al*[35]) share a 10 Mb IBD segment on chromosome 13 (ie, five markers), given the observed IBS sharing. The solid lines correspond to IBS sharing using five markers in linkage equilibrium with five equifrequent alleles (full black line), *vs* one common allele with frequency equal to 0.6 and four rare alleles with frequency equal to 0.1 (full gray line). The dashed lines correspond to IBS sharing using non-overlapping sets of three adjacent marker loci in complete linkage disequilibrium with four different haplotypes (black dashed line) *vs* non-overlapping sets of five adjacent marker loci with four haplotypes (gray dashed line); 10 000 simulations were used. The physical distance between adjacent markers was 2 Mb.

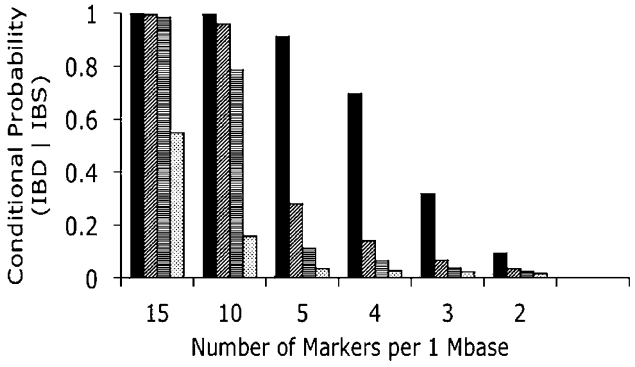

Figure 3 The probability that IBS sharing reflects sharing IBD for various marker spacing and informativity. The conditional probability that observed IBS sharing among third cousins reflects IBD sharing, given various marker spacing and marker-allele informativity based on 10 000 simulations over the human genome. The first two columns in each spacing category model microsatellite markers and the last two model single nucleotide polymorphisms. The microsatellite markers were assumed to harbor either five equally frequent alleles (full shading), or one frequent allele ($f=0.6$) and four rare alleles with equal frequency 0.1 (diagonal hatching). The single nucleotide polymorphisms were assumed to harbor either two equally frequent alleles (horizontal hatching), or one frequent allele ($f=0.8$) and the other allele with frequency 0.2 (dotted shading).

Haplotype-sharing probabilities for gene mapping studies

To showcase the utility of assessing the probabilities associated with chromosomal segment sharing in gene mapping studies, we used our program to assess the probabilistic significance of a reported finding of excess chromosome segment sharing among affected pedigree members in a study of Hirschsprung disease.[35] Puffenberger *et al*[35] ascertained a large, inbred, Mennonite pedigree (depicted in Figure 1 in Puffenberger *et al*[35]), exhibiting high incidence of Hirschsprung disease. Genealogical analysis of all members of the Mennonite pedigree identified a single common ancestral couple for all parents of the affected offspring. After searching for locations of the gene(s) responsible for Hirschsprung disease in this pedigree by genotyping three small multicase families and pursuing IBD allele-sharing analysis, the authors identified a 10 cM segment on chromosome 13 (corresponding to $\sim$10 Mb in this region) shared by all affected individuals. Follow-up LD analysis of 28 additional nuclear families confirmed the locus' association with disease. We obtained information about interlocus distance for the marker loci used by Puffenberger *et al*[35] through the MapViewer web interface of National Center for Biotechnology Information database.[36]

We estimated the probability that the pedigree members share the observed segment purely IBS. To do this, we had to make assumptions about allele frequencies and LD patterns among the loci. We also estimated the probability that this group of related pedigree members share such a segment IBD purely by chance, either anywhere on the chromosome or at the position in question. The results of the simulations suggest that the sharing observed among the pedigree members studied in the initial analysis has a probability of $\sim$0.15 in the region of interest.

We, then, considered the probability that the observed IBS allele sharing reflected IBD sharing. Figure 2 presents the conditional probability that the observed IBS sharing reflects IBD sharing based on 10 000 simulations assuming, first, that the marker alleles are in linkage equilibrium and have different allele frequencies, and, second, that non-overlapping sets of markers are in complete LD. To investigate how easily allele frequency information can influence the calculation of IBS sharing, and hence the conditional probability that IBS sharing reflects IBD sharing, we performed additional simulations with different assumptions about allele frequencies. The solid lines in Figure 2 depict the results of these studies, which involved evenly spaced markers with different allele frequencies. The markers were placed throughout the genome at 2 Mb intervals, and were assumed to harbor either five equally frequent alleles (black full line), or one frequent allele ($f=0.6$) and four rare alleles with an equal frequency of 0.1 (gray full line). In both studies, we assumed that the markers were in linkage equilibrium. The results suggest that marker informativity has large effect on IBS segment-sharing probabilities, as expected, since common alleles have high probability of being shared IBS.

We also performed simulations that investigated the effect of LD on the conditional probability that IBS sharing reflects IBD sharing. Figure 2 (dashed lines) presents the conditional probability that IBS sharing reflects IBD sharing using non-overlapping sets of three adjacent markers (spaced at 2 Mb intervals) in complete LD with four unique haplotypes (black dashed line), and non-overlapping sets of five adjacent markers with four unique haplotypes (gray dashed line). The dashed lines clearly show recurring 'drops' in conditional probability associated with chromosomal segments that extend over LD block boundaries. The conditional probability was assessed in overlapping chromosomal segments (ie, segments containing markers 1–5, 2–6, etc.).

The design of haplotype sharing-based mapping studies

Information on chromosomal segment sharing is also useful in the design stage of a haplotype-sharing study. To showcase this, we conducted simulations assuming a haplotype-sharing study that involved a pedigree with three affected third cousins. We estimated the distribution of the size of IBD segments shared by these third cousins in a specific region of chromosome 1. Our simulations suggest that 95% of IBD segments shared by three third cousins, purely by chance in the region of relevance, were longer than 1 Mb. Using further simulations, we addressed the question of how many markers would be necessary to detect these segments as IBD with a 20% false-positive rate given observed IBS sharing. Figure 3 displays the probability of sharing segments within marker sets spanning 1 Mb that reflect IBD sharing (and not purely IBS sharing) based on 10 000 simulation runs. It is evident that to obtain a false-positive rate $<20\%$ for the identification of IBD-shared segments (ie, the conditional probability of IBD given IBS must be higher than 0.8), one needs either highly informative microsatellite markers spaced every 200 kb or highly informative single nucleotide polymorphisms with interlocus distances at most 100 kb throughout the region.

The effect of crossover interference on chromosomal segment sharing

Finally, we explored the effect of crossover interference on the number and size of shared chromosomal segments. We devised a model of crossover interference that prevents the occurrence of nearby recombination events. We, then, computed the distribution of the size of chromosomal segments shared by two cousins on chromosome 1 using the crossover interference model as a function of the 'range of interference,' that is, the minimal distance separating the locations of two recombination events (for details see the Methods section). Our results suggest that modest range of interference does not seem to affect the average size or number of shared chromosomal segments significantly. However, when the range of interference is larger than approximately 50 Mb, fewer segments are shared, and the large shared segments tend to increase in size, resulting in increased variance in overall shared segment size. However, the proportion of the genome that is shared remains unchanged. Simulations that involved a more complex pedigree structure yielded similar results (data not shown).

The results generated under our interference model suggest that the previous results reported in this study, which were obtained under the assumption of no interference, would not change dramatically under an interference model (of at least the type we have devised). The distributions of the fraction of the genome that is shared would likely exhibit a slightly greater variance, but their means would remain the same. In analyses that assess sharing probabilities associated with chromosomal segments of a given size, the larger size of some shared segments would likely compensate for the smaller number of shared segments generated under the interference model.

Discussion

Many genetic research initiatives, especially genetic mapping initiatives, consider or require information about the sharing of genomic segments among individuals with a particular phenotype and known genealogical relationships. This information can involve the frequency of shared segments, the probability that segments of a certain size are shared, and the probability that individuals sharing alleles at a number of adjacent loci actually share those alleles because they are part of the chromosomal segment that the individuals in question have inherited from a common ancestor (ie, the segment is shared IBD and not just IBS). We have employed an extended, genome-wide gene-dropping simulation-based method to assess chromosomal segment sharing among individuals with arbitrary genealogical links. The method is intuitive and very flexible and can be used to quantify chromosome segment-sharing probabilities in a wide variety of contexts. In this study, we used information about actual or known population-derived recombination rates and, thus, were not constrained by assumptions inherent in theoretical recombination models (or mapping functions). Since empirical recombination maps correspond to the observed recombination counts obtained from meiotic events observed in a sample of individuals, they automatically incorporate information about genomic recombination hotspots and coldspots. In addition, the maps can easily be updated or replaced with new and more refined maps specifically tailored to particular organisms or populations.

In developing our simulation method, we wanted to exploit empirically derived information about phenomena

that impact the transmission of chromosomal segments from parents to offspring, thus making as few arbitrary assumptions as possible. Ideally, our method should employ not only empirical, region and sex-specific data about recombination rates, but also mutation and gene conversion information. Unfortunately, detailed mutation and gene conversion rates are not yet available on a genome-wide basis. In addition, it is quite likely that no more than a few mutation events can be expected to occur even in a large pedigree with many loci being studied. Since a single mutation affects very few genomic segments, its impact on the number or size distribution of shared segments throughout a chromosome or the entire genome would not be great. Therefore, we decided not to incorporate mutation information in the present study. Our method, however, can be expanded to incorporate such data once it becomes available. Similarly, one could expand our method to incorporate information about genotyping error associated with a particular genotyping technology.

Accurate estimates of allele frequencies may not always be available. Our results indicate (not shown) that slight misspecifications of allele frequencies do not affect the results substantially. However, when actual allele frequencies are not known, users should make assumptions concerning the range of likely allele frequencies based on available data, and compare sharing probabilities from several analyses that involve frequencies from that range.

There are many extensions and additional areas of application for the proposed procedure. For example, one could use the method to determine just how inbred a pair or group of individuals are in the absence of detailed genealogical information by comparing genome and segment-sharing probabilities calculated with and without assumptions about the degree of relatedness of the founders of the pedigree containing the individuals in question. This area of application is also important for putting any chromosomal segment sharing analysis into context, since initial analyses of available genealogical information may assume that all founders are unrelated. It may be important, therefore, to re-run the analysis with assumptions about the relationships between the founders (ie, hypothetical genealogical links between them). Ignoring the impact of undocumented inbreeding or false parenthood will negatively impact the results of any analysis of genome and segment sharing.[37] In cases when founders' relations are uncertain, the inbreeding pattern could be extracted from marker data by estimating individuals' inbreeding coefficients with a method like the one proposed by Leutenegger *et al*.[38] Although this information cannot be used to reconstruct the true genealogical history of the founders due to hidden consanguinity, the original pedigree combined with a genealogy reflecting the estimated inbreeding coefficient of the founders can then be used with the proposed simulation studies to obtain more accurate results that concern for example, determining the optimal marker density for haplotype-sharing studies.

Our method can also be extended to incorporate region-specific LD information in the calculation of IBS allele-sharing probabilities. This can be achieved by assigning entire haplotypes (instead of individual marker alleles) to founders according to the frequencies of these haplotypes. Alternatively, one can use LD data collected in more general populations from available public resources, for example, The International HapMap Project database.[39,40] This extension is especially useful in applications that involve pedigrees ascertained from a small, isolated, inbred population where one would expect the presence of strong LD.

Although strong evidence exists for positive crossover interference in the human genome,[21] there has been some discussion on appropriate mathematical models for it. Several studies[20–22] have explored gamma distribution-based models, in which distances between crossovers follow a gamma distribution, and concluded that these models fit samples of empirical data quite well. Such models are essentially a generalization of the model that we employed in the present study; instead of prohibiting all recombination events that would occur near the previous recombination event given the observed recombination rates, in the gamma models, such recombination events are prohibited with a given probability. Our simulation program can be adapted to model interference in this fashion. If we assume that the actual range of crossover interference that exists in the human genome does not extend far beyond 100 Mb, then, because our model rejects a greater number of recombination events than gamma models, our interference model can quantify the upper bound of the effect that interference has on genome sharing.

The source code and executables of the C++ program for conducting analysis and computing relevant probabilities along with the user manual are available for download at http://polymorphism.scripps.edu/people-files/genShare.tar.gz.

Acknowledgements

NJS is supported in part by The NHLBI Family Blood Pressure Program (FBPP; U01 HL064777-06); The NIA Longevity Consortium (U19 AG023122-01); the NIMH Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571-02); NIH R01s: HL074730-02 and HL070137-01; and Scripps Genomic Medicine.

References

1 Cavalli-Sforza LL, Bodner WF: *The Genetics of Human Population*. San Francisco: WH Freeman and Co, 1971.
2 Jacquard A: *The Genetic Structure of Populations*. New York: Springer-Verlag, 1974.

3 Risch N, Lange K: Application of a recombination model in calculating the variance of sib pair genetic identity. *Ann Hum Genet* 1979; **43**: 177–186.
4 Schork NJ, Thiel B, St Jean P: Linkage analysis, kinship, and the short-term evolution of chromosomes. *J Exp Zool* 1998; **282**: 133–149.
5 Genin E, Todorov AA, Clerget-Darpoux F: Optimization of genome search strategies for homozygosity mapping: influence of marker spacing on power and threshold criteria for identification of candidate regions. *Ann Hum Genet* 1998; **62**: 419–429.
6 Schaffer AA: Computing probabilities of homozygosity by descent. *Genet Epidemiol* 1999; **16**: 135–149.
7 Broman KL, Weber JL: Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am J Hum Genet* 1999; **65**: 1493–1500.
8 Clark AG: The size distribution of homozygous segments in the human genome. *Am J Hum Genet* 1999; **65**: 1493–1500.
9 Wiuf C, Hein J: The ancestry of a sample of sequences subject to recombination. *Genetics* 1999; **151**: 1217–1228.
10 McPeek MS, Strahs AL: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999; **65**: 858–875.
11 Stefanov VT: Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. *Genetics* 2000; **156**: 1403–1410.
12 Stefanov VT: Statistics on continuous IBD data: exact distribution evaluation for a pair of full (half)-sibs and a pair of a (great-) grandchild with a (great-) grandparent. *BMC Genet* 2002; **3**: 7.
13 Nolte IM, te Meerman GJ: The probability that similar haplotypes are identical by descent. *Ann Hum Genet* 2002; **66**: 195–209.
14 Chapman NH, Thompson EA: The effect of population history on the lengths of ancestral chromosome segments. *Genetics* 2002; **162**: 449–458.
15 Chapman NH, Thompson EA: A model for the length of tracts of identity by descent in finite random mating populations. *Theor Popul Biol* 2003; **64**: 141–150.
16 Leal SM, Yan K, Muller-Myhsok B: SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. *Hum Hered* 2005; **60**: 119.
17 Thompson EA: Monte Carlo in Genetic Analysis. Technical report no. 294, Department of Statistics, University of Washington 1995.
18 Thompson EA: Statistical Inferences from Genetic Data on Pedigrees. *NSF-CBMS Regional Conference Series in Probability and Statistics 2000*, Volume 6. OH: IMS, Beachwood.
19 George AW, Thompson EA: Multipoint linkage analyses for disease mapping in extended pedigrees: a Markov chain Monte Carlo approach. *Stat Sci* 2003; **18**: 515–531.
20 Zhao H, Speed TP, McPeek MS: Statistical analysis of crossover interference using the chi-square model. *Genetics* 1995; **139**: 1045–1056.
21 Broman KW, Weber JL: Characterization of human crossover interference. *Am J Hum Genet* 2000; **66**: 1911–1926.
22 Lin S, Cheng R, Wright FA: Genetic crossover interference in the human genome. *Ann Hum Genet* 2001; **65**: 79–93.
23 Yu A, Zhao C, Fan Y *et al*: Comparison of human genetic and sequence-based physical maps. *Nature* 2001; **15**: 951–953.
24 Sun F, Oliver-Bonet M, Liehr T *et al*: Human male recombination maps for individual chromosomes. *Am J Hum Genet* 2004; **74**: 521–531.
25 McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304**: 581–584.
26 Goldgar DE, Fain PR: Models of multilocus recombination: nonrandomness in chiasma number and crossover positions. *Am J Hum Genet* 1988; **43**: 38–45.
27 Goldgar DE, Fain PR, Kimberling WJ: Chiasma-based models of multilocus recombination: increased power for exclusion mapping and gene ordering. *Genomics* 1989; **5**: 283–290.
28 Karlin S, Liberman U: Theoretical recombination processes incorporating interference effects. *Theor Popul Biol* 1994; **46**: 198–231.
29 Windemuth C, Simianer H, Lien S: Fitting genetic mapping functions based on sperm typing: results for three chromosomal segments in cattle. *Anim Genet* 1998; **29**: 425–434.
30 MacCluer JW, Vandeberg JL, Read B, Ryder O: Pedigree analysis by computer simulation. *Zoo Biol* 1986; **5**: 147–160.
31 Ke X, Tapper W, Collins A: LDB2000: sequence-based integrated maps of the human genome. *Bioinformatics* 2001; **17**: 581–586.
32 LDB2000 sequence-based integrated maps of the human genome [http://cedar.genetics.soton.ac.uk/pub].
33 Centre d'Etude du Polymorphisme Humain panel, version 9 [http://www.cephb.fr/cephdb/].
34 Genatlas [http://bisance.citi2.fr/genatlas/].
35 Puffenberger EG: Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum Mol Genet* 1994; **3**: 1217–1225.
36 NCBI database [http://www.ncbi.nlm.nih.gov/mapview].
37 Libiger O, Schork NJ: Simulation-based homozygosity mapping with GAW14 COGA dataset on alcoholism. *BMC Genet* 2005; Suppl 1: S31.
38 Leutenegger AL, Prum B, Genin E *et al*: Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 2003; **73**: 516–523.
39 Altshuler D, Chakravarti A, Collins FS *et al*: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
40 The International HapMap Project database [http://www.hapmap.org].