ARTICLE

The Value of Molecular Haplotypes in a Family-Based Linkage Study

E. M. Gillanders, J. V. Pearson, A. J. M. Sorant, J. M. Trent, J. R. O'Connell, and J. E. Bailey-Wilson

Novel methods that could improve the power of conventional methods of gene discovery for complex diseases should be investigated. In a simulation study, we aimed to investigate the value of molecular haplotypes in the context of a family-based linkage study. The term "haplotype" (or "haploid genotype") refers to syntenic alleles inherited on a single chromosome, and we use the term "molecular haplotype" to refer to haplotypes that have been determined directly by use of a molecular technique such as long-range allele-specific polymerase chain reaction. In our study, we simulated genotype and phenotype data and then compared the powers of analyzing these data under the assumptions that various levels of information from molecular haplotypes were available. (This information was available because of the simulation procedure.) Several conclusions can be drawn. First, as expected, when genetic homogeneity is expected or when marker data are complete, it is not efficient to generate molecular haplotyping information. However, with levels of heterogeneity and missing data patterns typical of complex diseases, we observed a 23%–77% relative increase in the power to detect linkage in the presence of heterogeneity with heterogeneity LOD scores >3.0 when all individuals are molecularly haplotyped (compared with the power when only standard genotypes are used). Furthermore, our simulations indicate that most of the increase in power can be achieved by molecularly haplotyping a single individual in each family, thereby making molecular haplotyping a valuable strategy for increasing the power of gene mapping studies of complex diseases. Maximization of power, given an existing family set, can be particularly important for late-onset, often-fatal diseases such as cancer, for which informative families are difficult to collect.

In contrast to simple Mendelian disorders, susceptibility to common complex diseases such as cancer, type-2 diabetes, or Alzheimer disease is multifactorial and involves multiple genetic and environmental risk factors. Efforts to localize susceptibility genes involved in complex diseases have been limited by genetic heterogeneity, incomplete penetrance, phenocopies, and, frequently, late age at disease onset. Each of these factors can result in a significant reduction in statistical power for any individual gene-mapping study. Thus, novel methods that could improve the power of traditional methods of gene discovery for complex diseases should be examined. Complex diseases include most common diseases of adult life, and they account for most human morbidity and mortality. Therefore, improvements in the methods used to decipher their genetic etiologies should be of paramount importance.

Reconstruction of haplotypes has proven critical to several studies that have succeeded in identifying genetic factors involved in complex-trait susceptibility.[1–4] Haplotypes provide additional information to both linkage and linkage disequilibrium (LD) studies and therefore may facilitate the mapping of a disease gene by allowing a more precise localization of the gene within a chromosomal region initially identified by linkage analysis. Candidate regions for complex diseases, initially identified by genomewide linkage scans, can often be prohibitively large (20–30 cM). These regions can contain upwards of 100 genes, which requires further narrowing of the candidate interval before positional cloning efforts.

Recent advances in molecular technologies and the availability of the human genome sequence have revolutionized researchers' ability to catalogue human genetic variation. However, reconstruction of haplotypes from conventional genotypes in diploid organisms such as humans can be complicated, since the parental origins of the two alleles of each genotype are not directly observable. There are three principle haplotyping approaches: (1) statistical estimation, (2) inference from family data, and (3) empirical (or "direct") molecular haplotyping. The reliability of statistical methods in reconstruction of haplotypes depends on the number of markers, allele frequencies, fraction of missing data, genotyping error rate, and LD between markers.[5–7] Inferring phase from family data can be limited by uninformative or missing genotypes. In addition, late age at onset for many complex diseases can preclude collection of DNA samples from previous generations, thereby further limiting strategies to reconstruct haplotypes with use of family data. In contrast, molecular haplotyping methods are empirical and are not dependent on statistical assumptions or estimation.

Two popular molecular haplotyping methods include (1) long-range, allele-specific PCR (AS-PCR)[8–10] and (2) diploid-to-haploid methods, such as conversion.[11] Crawford and Nickerson[12] describe these methods in detail. In brief,

From the Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore (E.M.G.; A.J.M.S.; J.E.B.-W.); Translational Genomics (TGen) Research Institute, Phoenix (J.V.P.; J.M.T.); and University of Maryland, Baltimore (J.R.O.)

Received April 14, 2006; accepted for publication June 12, 2006; electronically published June 28, 2006.

Address for correspondence and reprints: Dr. E. M. Gillanders, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, 333 Cassell Drive, Suite 1200, Baltimore, MD 21224. E-mail: lgilland@mail.nih.gov

Am. J. Hum. Genet. 2006;79:458–468. 0002-9297/2006/7903-0009$15.00

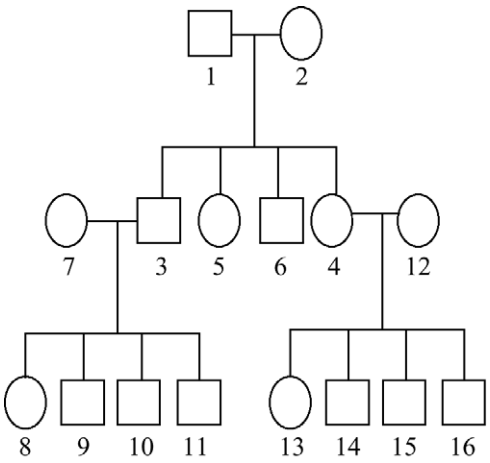

Figure 1. Pedigree structure used in simulation analyses. Each pedigree was required to have three affected individuals for ascertainment. Which three individuals were affected was somewhat limited, to minimize the inclusion of families uninformative for linkage analysis.

these molecular techniques unambiguously reconstruct haplotypes in the following manner. AS-PCR involves selective PCR amplification of one of the two chromosomes at a given heterozygous locus. This is frequently done by designing PCR primers that will match (or mismatch) one allele at the 3′ end of the primer. By use of long-range PCR methods, a molecular haplotype of up to 40 kb can be determined. The conversion method entails generation of mouse-human somatic cell hybrids, which retain only a subset of human chromosomes. Once hybrids that are monosomic for the chromosome of interest are identified, haplotypes can be reconstructed with conventional genotyping of the haploid cells. In short, both methods provide unequivocal molecular haplotypes. Several studies have provided evidence of the value of molecular haplotyping in the context of LD studies.[13–15]

In this study, we used simulations to compare the power of using various levels of molecular haplotypes (available because of the simulation procedure) with the power of using standard genotyping in the context of a family-based linkage study. To clarify, we are not assuming any molecular haplotyping method in particular; we are simply assuming that we have molecularly derived haplotypes available. These results will need to be considered within the context of considerably increased laboratory expenditure. Current molecular haplotyping methods are fairly limited and not particularly well suited for high-throughput work; however, positive results might motivate new molecular or statistical approaches that could more easily capitalize on the benefits of molecular haplotyping information.

Material and Methods

Using the Genometric Analysis Simulation Program (GASP),[16] we simulated qualitative trait and marker data for the three-generation pedigree structure shown in figure 1. The simulated qualitative trait was due to a single locus with an autosomal dominant mode of inheritance and a disease allele (D) frequency equal to 0.01. Individuals with both the DD and Dd genotypes had an 80% probability of developing the disease. Of individuals with a normal genotype (dd), 4% developed the trait. All individuals within the pedigree were considered to be beyond the age of risk.

Families were ascertained (i.e., selected from randomly simulated families to be in the analyzed sample) on the basis of having a minimum of three affected members who were at least minimally informative for linkage analysis. Specifically, families were excluded if the three affected members were (1) all founders (fig. 2*A*), (2) a parent-parent-offspring trio (fig. 2*B*), (3) a parent-off-

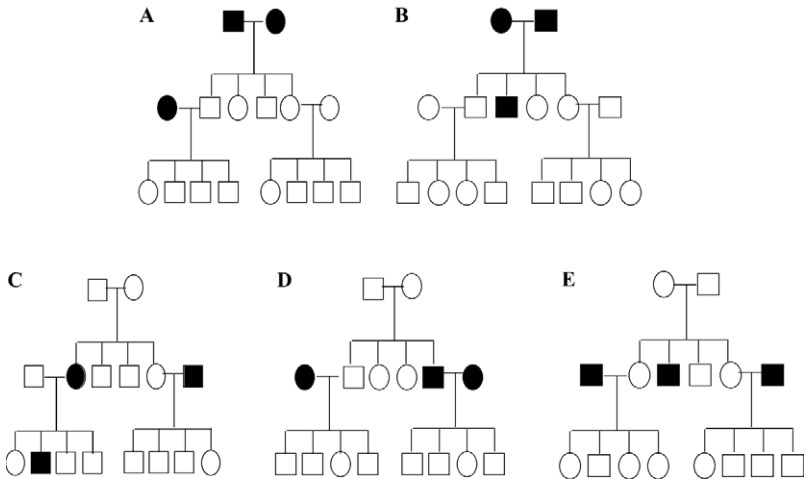

Figure 2. Examples of pedigrees excluded from simulations. Families were not included if the three affected members were all founders (*A*), a parent-parent-offspring trio (*B*), a trio of parent, offspring, and a founder who is not a grandparent (*C*), a founder-founder-spouse trio (*D*), or a trio of individuals 7, 12, and either 5 or 6 (*E*) (individual numbering as in fig. 1).

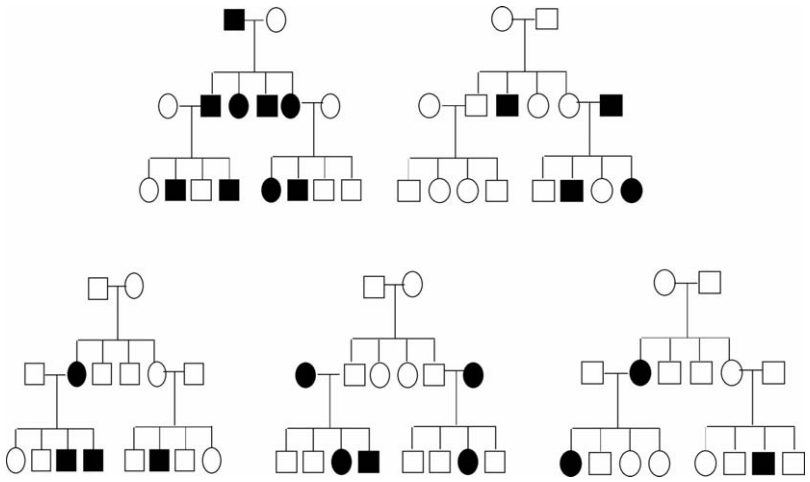

Figure 3. Examples of specific pedigrees included in simulation analyses

spring pair and a founder who is not a grandparent (fig. 2*C*), (4) a founder-founder-spouse trio (fig. 2*D*), or (5) a trio comprising individuals 7, 12, and either 5 or 6 (fig. 2*E*). This ascertainment procedure was designed to mimic the real procedures used in linkage studies of qualitative traits. Simulations of families for each replicate continued until there were 100 pedigrees that met our ascertainment criteria. Examples of pedigrees included in our simulation are shown in figure 3.

GASP was used to simulate genotype data for eight STR markers, including four STR genomewide-scan (GWS) markers that were 10 cM apart and four STR fine-mapping (FM) markers that were 1 cM apart. Each simulated STR marker had five equally frequent alleles. In separate simulations, we generated genotype data for four SNP markers. These SNP GWS markers were 1 cM apart and had a minor-allele frequency of 0.40. The genetic marker loci were all assumed to be in linkage equilibrium. Furthermore, because the marker data were simulated, the haplotypes were known with certainty for all individuals in the data set. Therefore, unlike a real linkage study (where only the genotype at each marker is known), we were able to write out the simulated marker data showing which allele was maternally or paternally derived for all loci in all individuals. This is the information that would be available if that person were molecularly haplotyped. We then included various levels of these simulated molecular haplotypes in the linkage analyses, as described below.

All simulations are outlined in figures 4–7. In each simulated family, the trait could be either linked or unlinked to the analyzed marker set. We simulated three different levels of genetic heterogeneity in which the percentage of linked families varied: (1) none of the families linked to the marker set (H0), (2) 100% of the families linked to the marker set (H1), and (3) 25% of the families linked to the marker set (H2). The H0 level of heterogeneity was simulated, to confirm that the inclusion of molecular haplotypes did not cause increases in type I error rates. The H1 and H2 levels of heterogeneity were used to determine the extent to which the inclusion of molecular haplotype information into the linkage analysis improved power.

For each level of heterogeneity (H0, H1, and H2), at least nine different models of data availability were considered (fig. 4). Specifically, for each replicate, we considered three different levels of haplotyping information: (1) no molecular haplotypes were included in the linkage analysis (only individual genotypes were used) (P0), (2) simulated molecular haplotypes of affected individuals were included in the linkage analysis (only genotypes of unaffected individuals were used) (P1), and (3) simulated molecular haplotypes from all genotyped individuals were included in the linkage analysis (P2). These different levels of simulated molecular haplotyping were used to determine whether all genotyped persons in the pedigree needed to be molecularly haplotyped to increase linkage power or whether linkage power gains

Table 1. Type I Error of Four STR FM Markers (1 cM Apart), with 0% Families Linked (H0)

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD >1.0	>2.0	>3.0
1	All members genotyped (P0)	No marker data missing (M0)	1.8	.2	.0
2	All members genotyped (P0)	Marker data missing for top generation (M1)	1.6	.1	.0
3	All members genotyped (P0)	Marker data missing for top generation and 50% of middle generation (M2)	1.9	.2	.0
4	Affected individuals haplotyped (P1)	No marker data missing (M0)	1.8	.2	.0
5	Affected individuals haplotyped (P1)	Marker data missing for top generation (M1)	1.8	.3	.0
6	Affected individuals haplotyped (P1)	Marker data missing for top generation and 50% of middle generation (M2)	1.2	.2	.0
7	All members haplotyped (P2)	No marker data missing (M0)	1.8	.2	.0
8	All members haplotyped (P2)	Marker data missing for top generation (M1)	1.8	.3	.0
9	All members haplotyped (P2)	Marker data missing for top generation and 50% of middle generation (M2)	1.9	.1	.1

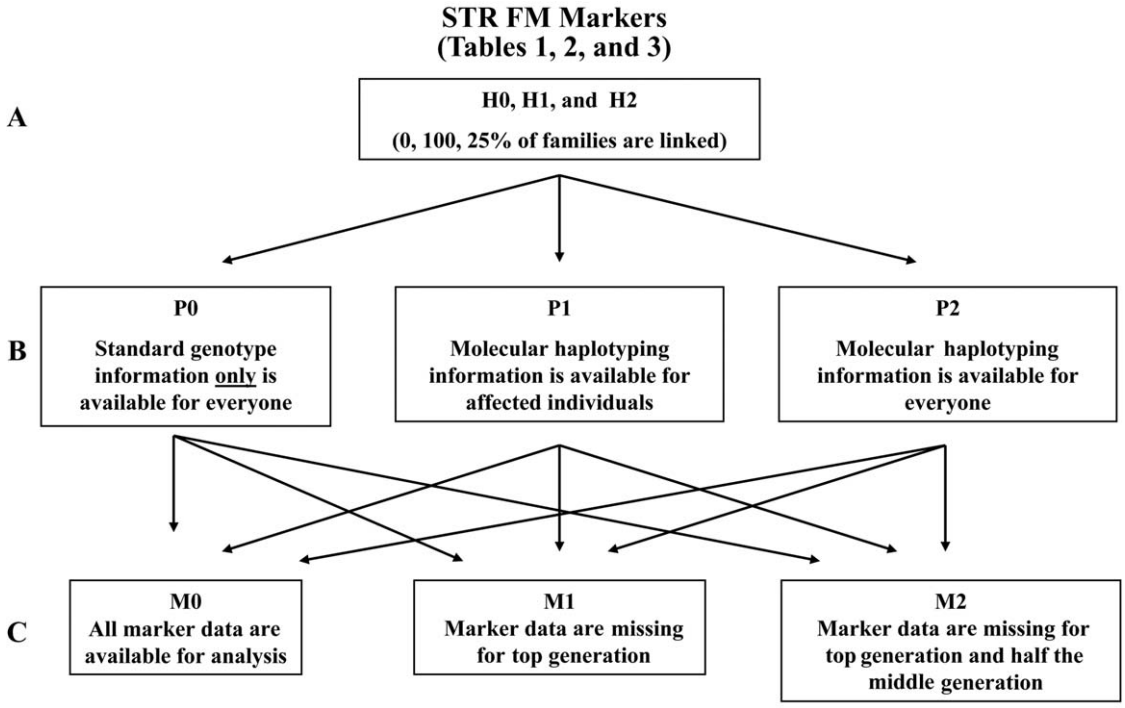

Figure 4. Outline of nine STR FM analyses for all three levels of genetic heterogeneity (H0, H1, and H2). (Results are shown in tables 1–3). For each level of genetic heterogeneity (*A*), we considered three levels of molecular haplotyping information (P0, P1, and P2) (*B*), and, for each level of molecular haplotyping information, we considered three levels of missing data (M0, M1, and M2) (*C*).

could be obtained by performing molecular haplotyping on a smaller proportion of the family members.

For each level of molecular haplotyping information (P0, P1, and P2), we considered three levels of missing marker data (i.e., different patterns of ungenotyped family members): (1) no missing data (marker data available for everyone) (M0), (2) marker data missing for individuals in our top generation of our simulated pedigree (individuals 1 and 2) (M1), and (3) marker data missing for individuals in our top generation as well as for 50% of individuals in our second generation (a randomly chosen half of founders and a randomly chosen half of nonfounders) (M2). We used these different missing data rates to evaluate whether including molecular haplotype data in the linkage analysis always increased power or whether it increased power only in the presence of incomplete family genotype data. When marker data were missing for an individual, that person's genotypes (at all marker loci) and simulated molecular haplotype were all treated as unknown in the linkage analysis. For each of these 27 models (nine basic data availability combinations times three levels of genetic heterogeneity), a set of four STR FM (1-cM spacing) markers surrounding the trait locus was analyzed (fig. 4).

For heterogeneity model H2 only, these four FM STR markers were also analyzed for two more levels of haplotyping information: simulated molecular haplotyping for one randomly chosen member of our middle generation (P3) and simulated molecular haplotyping for one randomly chosen member of our bottom generation (P4), for each level of missing data (M1, M2, and M3) (fig. 5). Once again, these two levels of simulated molecular haplotyping were used to determine whether all genotyped persons in the pedigree needed to be molecularly haplotyped to increase linkage power or whether linkage power gains could be obtained by performing molecular haplotyping on only a small proportion

Table 2. Power of Four STR FM Markers (1 cM Apart), with 100% of Families Linked (H1)

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD		
			>1.0	>2.0	>3.0
1	All members genotyped (P0)	No marker data missing (M0)	100.0	100.0	100.0
2	All members genotyped (P0)	Marker data missing for top generation (M1)	100.0	100.0	100.0
3	All members genotyped (P0)	Marker data missing for top generation and 50% of middle generation (M2)	100.0	100.0	100.0
4	Affected individuals haplotyped (P1)	No marker data missing (M0)	100.0	100.0	100.0
5	Affected individuals haplotyped (P1)	Marker data missing for top generation (M1)	100.0	100.0	100.0
6	Affected individuals haplotyped (P1)	Marker data missing for top generation and 50% of middle generation (M2)	100.0	100.0	100.0
7	All members haplotyped (P2)	No marker data missing (M0)	100.0	100.0	100.0
8	All members haplotyped (P2)	Marker data missing for top generation (M1)	100.0	100.0	100.0
9	All members haplotyped (P2)	Marker data missing for top generation and 50% of middle generation (M2)	100.0	100.0	100.0

Table 3. Power of Four STR FM Markers (1 cM Apart), with 25% of Families Linked (H2)

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD >1.0	>2.0	>3.0
1	All members genotyped (P0)	No marker data missing (M0)	98.6	94.5	87.0
2	All members genotyped (P0)	Marker data missing for top generation (M1)	97.3	88.8	74.9
3	All members genotyped (P0)	Marker data missing for top generation and 50% of middle generation (M2)	95.0	81.0	63.1
4	Affected individuals haplotyped (P1)	No marker data missing (M0)	98.6	94.5	87.0
5	Affected individuals haplotyped (P1)	Marker data missing for top generation (M1)	98.7	94.3	86.6
6	Affected individuals haplotyped (P1)	Marker data missing for top generation and 50% of middle generation (M2)	97.5	89.3	74.4
7	All members haplotyped (P2)	No marker data missing (M0)	98.6	94.5	87.0
8	All members haplotyped (P2)	Marker data missing for top generation (M1)	98.7	94.6	87.0
9	All members haplotyped (P2)	Marker data missing for top generation and 50% of middle generation (M2)	98.1	90.4	77.4

of the family members. Again, for heterogeneity model H2 only and the basic nine data-availability combinations, we also analyzed two additional marker sets: four GWS STR markers 10 cM apart (fig. 6) and four GWS SNP markers 1 cM apart (fig. 7). For each described scenario, we analyzed 1,000 replicates.

We used version 2 of the VITESSE program[17,18] to perform multipoint linkage analyses. VITESSE is a linkage-analysis program that uses the Elston-Stewart algorithm to compute the likelihood of pedigree data. VITESSE uses a novel set-recoding scheme to reduce the number of genotypes required in the likelihood calculation, thereby improving the computational performance. The program accepts a wide variety of special input formats, including phased genotype data. Standard genotype-input format does not distinguish the parental source of alleles in heterozygous genotypes. VITESSE permits a vertical bar delimiter (|) between alleles to set the paternal source of the allele. For example, an input genotype of 1|2 in the pedigree file is read by VITESSE as specifying the alleles 1 and 2 as paternally and maternally inherited, respectively. Moreover, this input format can be used to specify paternally inherited *n*-locus haplotypes at *n* loci by specifying phased genotypes at each of the *n* loci. This is how our simulated molecular haplotypes were specified in the linkage analysis. Multipoint linkage analyses were performed under the assumption that the genetic model is the same as the generating one.

In multipoint likelihood-based linkage analysis, the likelihood of the marker data conceptually is a sum over all possible haplotype configurations, and phase information for one or more individuals in the family reduces the number of configurations. When VITESSE reads in molecularly determined haplotypes as described above, a reduction in the number of possible haplotype configurations in the family is done at the time of genotype input. Thereafter, VITESSE performs the standard likelihood calculations over the remaining possible haplotype configurations in the family. It is this reduction in the uncertainty about the true haplotype configuration in the family that is expected to increase power to detect linkage when molecular haplotypes are included in the linkage analysis.

Heterogeneity LOD–score (HLOD) calculations to test for linkage in the presence of genetic heterogeneity were performed using the admixture test.[19] Power to detect linkage (or, in the completely unlinked situation, type I error rate) was measured as the percentage of 1,000 replicates that reached an HLOD of at least 1.0, at least 2.0, and at least 3.0.

Analyses were run on the TGen Research Institute IBM 1350 computational cluster. The cluster had 512 IBM X Series computational nodes, each with two 2.4 GHz Intel Xeon processors, for a total of 1,024 central processing units. Each node had 2 GB of RAM and Gigabit Ethernet network connections. The operating system was RedHat Enterprise Linux 3.0, PBSPro was used to handle job scheduling, and the IBM CSM (Cluster System Management) was used to monitor and maintain the cluster. Of the nodes, 128 are equipped with low-latency, high-throughput Myrinet interconnects. These nodes also have an extra 2 GB of RAM per node to allow for memory-intensive computations. PBSPro manages these resources and allows jobs to be run exclusively on these Myrinet nodes. All cluster nodes have access to a shared parallel 1 TB file system (IBM GPFS), which allows each individual node to read and write to the same data files simultaneously across the cluster. The GPFS file system uses IBM FAStT SAN (storage area network) storage units, providing high performance and high reliability.

Results

Four STR FM (1-cM Spacing) Marker Analyses

Summary results for the four STR FM marker multipoint analyses are presented in tables 1, 2, and 3 for genetic heterogeneity models H0, H1, and H2, respectively. In all

Table 4. Power of Four STR FM Markers (1 cM Apart), with 25% of Families Linked (H2) and with Additional Molecular Haplotyping Levels

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD >1.0	>2.0	>3.0
10	One individual haplotyped (middle generation) (P3)	No marker data missing (M0)	98.6	94.5	87.0
11	One individual haplotyped (middle generation) (P3)	Marker data missing for top generation (M1)	98.5	92.1	82.6
12	One individual haplotyped (middle generation) (P3)	Marker data missing for top generation and 50% of middle generation (M2)	97.0	87.1	71.7
13	One individual haplotyped (bottom generation) (P4)	No marker data missing (M0)	98.6	94.5	87.0
14	One individual haplotyped (bottom generation) (P4)	Marker data missing for top generation (M1)	97.3	88.8	74.9
15	One individual haplotyped (bottom generation) (P4)	Marker data missing for top generation and 50% of middle generation (M2)	94.7	81.1	62.8

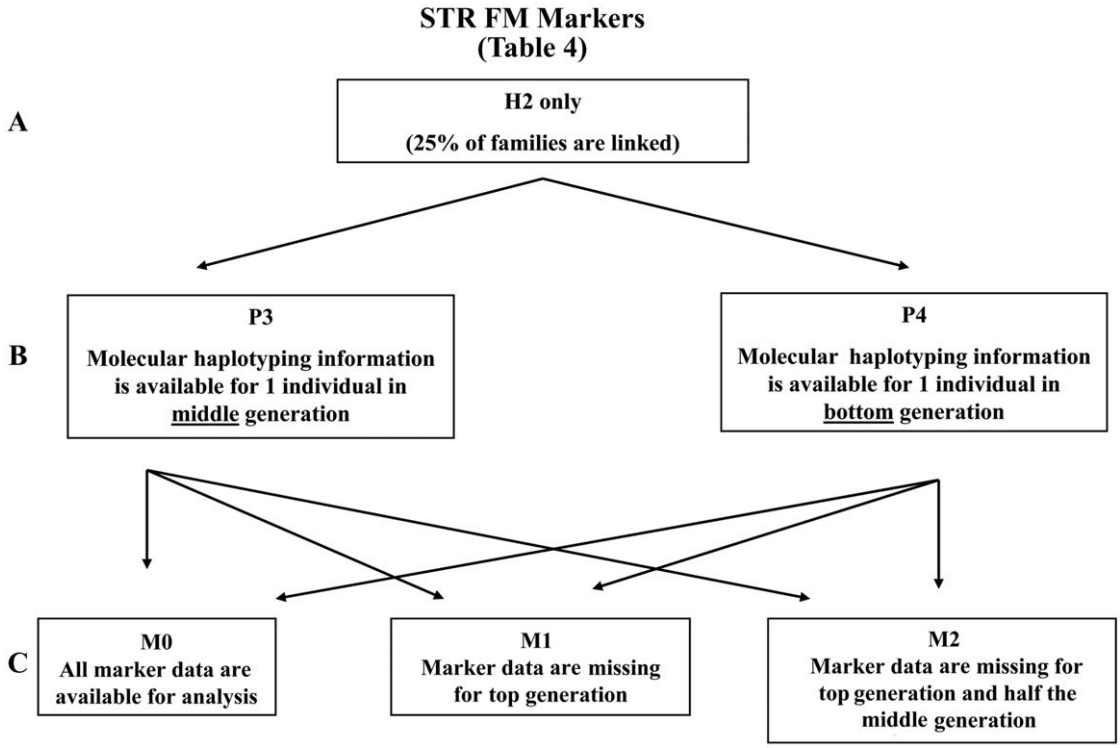

Figure 5. Outline of additional STR FM analyses (H2 only). For the H2 level of heterogeneity only (*A*), we considered two additional levels of molecular haplotyping information (*B*), each with molecular haplotyping information available for a single individual. For each of these two levels of molecular haplotyping information, we considered three levels of missing data (M0, M1, and M2) (*C*). Results are shown in table 4.

simulations, models 1–3 assume no one has been molecularly haplotyped (P0); thus, only genotype data were included in the linkage analysis. Models 4–6 were analyzed under the assumption that all genotyped affected individuals had been molecularly haplotyped (P1) (with use of some molecular method), so that simulated molecular haplotypes for the affected individuals as well as genotypes for unaffected individuals were used in the analysis. Finally, models 7–9 assume everyone has been molecularly haplotyped (P2), so that simulated molecular haplotypes were included in the analysis of all genotyped family members. In table 4 (heterogeneity model H2 only), haplotyping schemes P3 and P4 (one family member molecularly haplotyped) are presented. In these models (10–15) the simulated molecular haplotype for one individual and genotypes for the remaining family members were used in the analysis. For each phase information level, three missing marker data options (M0, M1, and M2) were considered.

Table 1 presents a summary of HLOD results for the replicates in which none of the simulated families have their trait locus linked to the markers tested (H0). In this situation, any linkage detected was a type I error. The HLOD results across different molecular haplotyping and missing marker–data options were similar and close to expected values. For each model, <2% of HLODs were >1.0. Less than 0.5% of HLODs were >2.0, and only a single replicate had an HLOD >3.0.

Table 2 presents a summary of HLODs when 100% of the ascertained families are linked to the simulated marker set (H1) and the same four tightly spaced markers are included in the multipoint analyses. All 1,000 replicates, for each of the nine different data availability models, had LOD (data not shown) and HLODs >3.0.

Table 3 gives a summary of HLODs when 25% of the families are linked to the simulated marker set (H2) and the same four STR FM markers are included in the multipoint analyses. When no one was missing (either genotypes or haplotypes) marker data (M0) (models 1, 4, and 7), there was no difference in results across levels of molecular haplotyping information included in the linkage analysis. When our top generation (individuals 1 and 2) was missing marker data (M1) (models 2, 4, and 8), there was a small relative increase in power to detect linkage with HLODS >1.0, >2.0, and >3.0 when direct haplotyping information was available for everyone (relative increases of 1%, 7%, and 16%, respectively). Missing marker–data pattern M2 (models 3, 6, and 9) assumes marker data are missing for founders in our top generation as well as for 50% of individuals in our middle generation. With this level of missing data, there was a larger increase in power

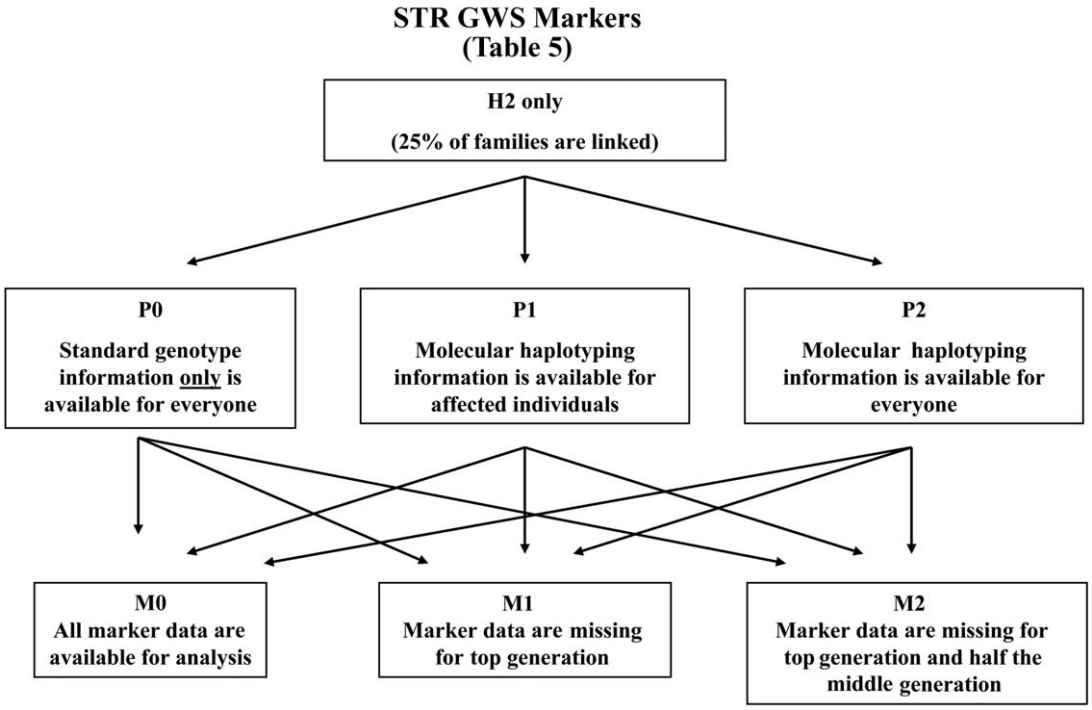

Figure 6. Outline of additional STR GWS analyses (H2 only). For the H2 level of heterogeneity only, we also analyzed four STR markers at GWS density (markers 10 cM apart). Results are shown in table 5.

to detect HLODs >1.0, >2.0, and >3.0 when all individuals are haplotyped (relative increases of 3%, 12%, and 23%, respectively).

When all individuals were haplotyped (P2), the power was only slightly greater than when only the affected individuals were haplotyped (P1). For example, with the more-extreme missing data pattern (M2), the power to detect linkage with HLODs >3.0 increased notably from 63.1% when individuals are genotyped (P0) (model 3) to 74.4% when molecular haplotypes from affected individuals were included in the linkage analysis (P1) (model 6) but only slightly more, to 77.0%, when everyone's molecular haplotypes were included in the analysis (P2) (model 9). In these simulated pedigrees, having molecular haplotyping information for everyone (compared with only for affected individuals) translates into determining molecular haplotype information for up to 13 additional unaffected individuals (fig. 3). When only one individual was haplotyped (table 4), the power to detect linkage with HLODs >3.0 increased from 63.1% when only individuals were genotyped to 71.7% when the single molecularly haplotyped individual was in our middle generation (P3) (model 3 in table 3, compared with model 12 in table 4). In contrast, if the molecularly haplotyped individual was in our bottom generation (P4), then the power was virtually the same as when only individuals were genotyped, for all HLOD criteria considered (model 3 in table 3 compared with model 15 in table 4).

Four STR GWS (10 cM Apart) Marker Analyses

Table 5 summarizes power results for multipoint analyses of a four STR marker set with GWS spacing for simulations in which 25% of the families are linked to the simulated marker set (H2). With a less dense set of STR markers, power was lower for all options considered than with the FM marker set (as expected), but incorporation of molecular haplotypes into the linkage analysis provided a small relative increase in the power to detect linkage with an HLOD of $\geqslant$3.0, even when no marker data were missing. When some marker data were missing, the improvement in power with the inclusion of molecular haplotypes was more striking, especially when HLOD was >3.0. For the M1 and M2 missing-data patterns, the relative increase was 22% and 43%, respectively, in the power with HLODs >3.0 (models 8 and 9). Again, most of the increase in power was gained by including the simulated molecular haplotypes from only the affected individuals.

Four SNP GWS (1 cM Apart) Marker Analyses

Table 6 provides a summary of the results of analyses of four SNP GWS markers when ~25% of the families were linked to the simulated trait (H2). With use of less informative SNP markers, there was again a small increase in the power to detect linkage with an HLOD of $\geqslant$3.0 even when no marker data were missing and molecular haplotype information was incorporated. When marker data

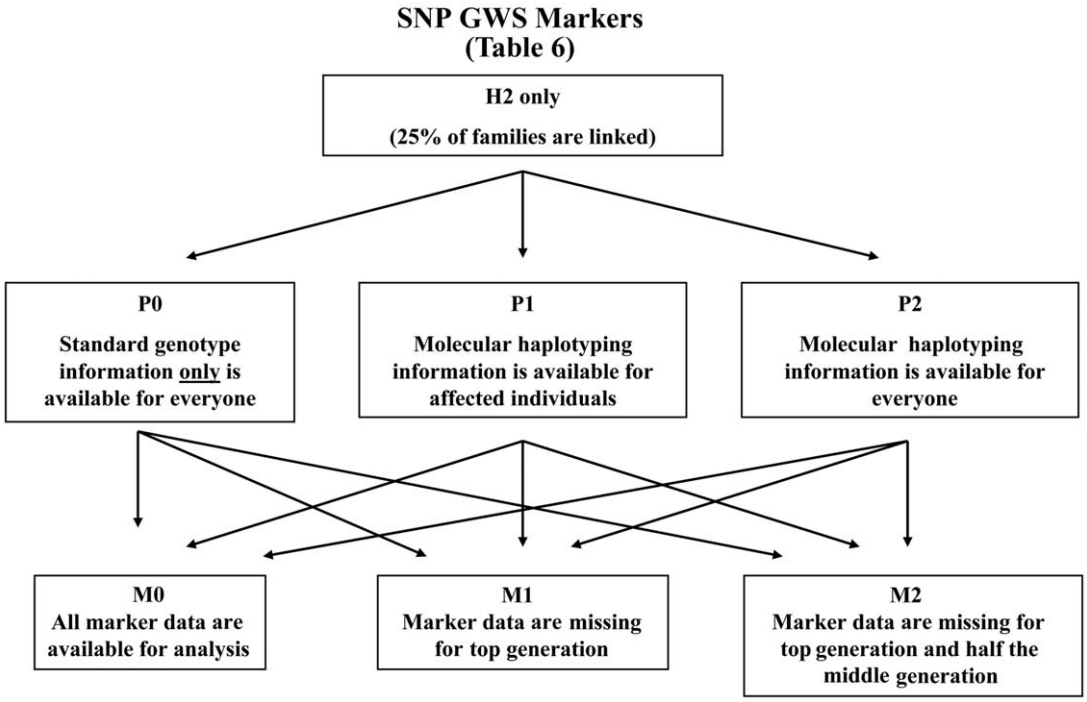

Figure 7. Outline of additional SNP GWS analyses (H2 only). For the H2 level of heterogeneity only, we also analyzed four SNP markers at GWS density (markers 1 cM apart). Results are shown in table 6.

were missing, there was a substantial increase in the power to detect linkage with an HLOD >3.0 when simulated molecular haplotypes for all nonmissing individuals were used in the analysis, compared with when only genotypes were used. When marker data were missing only for our top generation (M1), the power to detect linkage with an HLOD >3.0 increased 33% when simulated molecular haplotyping information for all nonmissing individuals was used in the analysis (model 8). When marker data were missing for our top generation and 50% of our middle generation (M2), power with an HLOD >3.0 increased 77% when simulated molecular haplotypes on all individuals were included in the analysis (model 9).

Discussion

Efforts to localize susceptibility genes involved in complex diseases such as cancer have been hindered by genetic heterogeneity, incomplete penetrance, phenocopies, and variable age at disease onset. Each of these factors results in a substantial loss of statistical power for any gene-mapping study. The goal of the present study was to explore the application of molecular haplotyping to linkage studies and the concomitant effects on power and type I error in a family-based linkage study. Haplotype reconstruction can increase the information available for linkage and LD studies, thus improving their power to identify genetic factors involved in complex disease susceptibility.

However, computational methods of reconstructing haplotypes can be limited as the number of markers increases or LD between markers decreases.[20,21] Haplotype inference from family data can be limited by uninformative or missing genotypes, particularly when the age at disease onset is late. Furthermore, most software packages used to infer haplotypes from pedigree data (such as GENEHUNTER[22] and SIMWALK2[23]) assume linkage equilibrium among markers, which may not be appropriate for tightly spaced markers.[24] In the current study, we used simulations to compare the power of using direct molecular haplotyping, to allow inclusion of molecular haplotypes in linkage analysis, with that obtained when only standard genotyping is available in the context of a family-based FM linkage study.

We used GASP to simulate a qualitative trait and marker data for extended pedigrees. Our intention was to make this simulated qualitative trait representative of a complex disease. In this regard, disease penetrance was modeled as incomplete, and there was a high rate of phenocopies, which is typical of complex diseases such as melanoma or prostate cancer. In addition, we also simulated three different levels of genetic heterogeneity. Finally, we considered three different levels of missing-marker data, the last of which could be considered typical of a disease of late onset age.

We evaluated the effect of both (1) increased phase information due to inclusion of molecularly determined haplotypes in the linkage analysis and (2) decreased in-

Table 5. Power of Four STR GWS Markers (10 cM Apart), with 25% of Families Linked (H2)

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD >1.0	>2.0	>3.0
1	All members genotyped (P0)	No marker data missing (M0)	98.3	92.8	81.9
2	All members genotyped (P0)	Marker data missing for top generation (M1)	96.0	84.4	67.3
3	All members genotyped (P0)	Marker data missing for top generation and 50% of middle generation (M2)	91.3	73.5	50.2
4	Affected individuals haplotyped (P1)	No marker data missing (M0)	98.5	93.0	83.0
5	Affected individuals haplotyped (P1)	Marker data missing for top generation (M1)	98.3	92.1	81.9
6	Affected individuals haplotyped (P1)	Marker data missing for top generation and 50% of middle generation (M2)	95.0	84.8	68.2
7	All members haplotyped (P2)	No marker data missing (M0)	98.6	93.3	82.9
8	All members haplotyped (P2)	Marker data missing for top generation (M1)	98.5	92.6	81.8
9	All members haplotyped (P2)	Marker data missing for top generation and 50% of middle generation (M2)	96.4	86.4	71.8

formation due to missing data on the type I error rate by simulating 1,000 replicates with the qualitative trait unlinked to the marker loci in 100% of ascertained families (H0). Table 1 shows the results of analysis of four STR FM markers for nine molecular haplotype and missing data scenarios. We conclude that the incorporation of molecular haplotyping information does not increase the type I error rate in an extended-pedigree FM linkage study. As we expected, results were similar across the various levels of molecular haplotyping information included in the linkage analysis. In addition, our results show that, in extended pedigrees, missing marker data also do not appear to inflate the type I error rate. We considered three different missing marker data patterns, including one in which genotypes or haplotypes were missing for founders in our top generation as well as for 50% of individuals in our middle generation. This level of missing data is characteristic of complex disorders with late onset age, because parents of affected individuals are frequently deceased. Table 1 shows no increase in the type I error rate with increasing missing-marker data.

When the simulated qualitative trait was linked to the simulated marker loci in all pedigrees (i.e., when there was linkage homogeneity) (H1) and four STR FM markers were included in the multipoint analyses, power to detect linkage was complete. In all such simulated data sets, regardless of the amount of molecular haplotyping information available or the level of missing-marker data, each replicate had LOD scores (not shown) and HLODs >3.0 (table 2). Given the extremely high level of power when there is no genetic heterogeneity, it is not surprising that the presence or absence of molecular haplotyping information had no effect on these results. This genetic homogeneity would be more characteristic of a simple Mendelian disorder than of a complex disease. As expected, this would not be the type of study that would benefit from the incorporation of molecular haplotyping information. Thus, for simple Mendelian disorders, one could not justify the increased cost of molecular haplotyping methodologies.

Finally, we simulated the situation in which 25% of the families were linked to the simulated marker set. First, we included four highly informative, closely spaced (1 cM apart) FM STR markers in our analyses and saw that, in the absence of missing-marker data, there is no benefit in including molecular haplotyping information. However, for a more realistic missing-data pattern for a complex disease (M1 or M2), inclusion of directly measured molecular haplotypes can increase the power to detect linkage with an HLOD of 3.0 by as much as 23% (table 3). Of course, these results must be balanced against the increased cost of molecular haplotyping methodologies. Because molecular haplotyping of everyone would be prohibitively expensive, we also considered intermediate schemes in which only certain individuals would be haplotyped. All of our simulations show that most of the improvement in power gained by incorporating molecular haplotyping information can be captured simply by directly haplotyping a small number of individuals. Specifically, we can capture 60% of this increase in power simply by molecular haplotyping a carefully selected single individual (table 4).

Continuing with this level of genetic heterogeneity (25% of families linked), which is more characteristic of a complex disease, we also analyzed a set of STR markers at a GWS density of 10 cM apart. As expected, the value of molecular haplotypes is even greater with a less dense set of markers (table 5). With missing-data patterns typical of a complex disease, there was a 22%–43% relative increase in the power to detect linkage with an HLOD >3.0. In the absence of a biologically relevant candidate gene, many gene-mapping studies of complex diseases begin with a global GWS for linkage. Our results show that the incorporation of molecular haplotyping information can substantially increase the power of these studies. Current trends for genomewide linkage studies favor genotyping thousands of SNP markers instead of hundreds of STR markers. Our simulations suggest that the value of direct haplotyping information may be even greater in such SNP-based studies, particularly for diseases with late onset, when substantial marker data will be missing (table 6). With substantial missing-marker data, the power to detect linkage with an HLOD >3.0 increased by 77% with the incorporation of haplotype information. Several studies have now reported an increase in linkage power with use

Table 6. Power of Four SNP GWS Markers (1 cM Apart), with 25% Families Linked (H2)

Model	Molecular Haplotyping Information	Options for Missing Marker Data	Percentage of 1,000 Replicates with HLOD >1.0	>2.0	>3.0
1	All members genotyped (P0)	No marker data missing (M0)	97.0	88.0	71.6
2	All members genotyped (P0)	Marker data missing for top generation (M1)	92.9	73.8	55.3
3	All members genotyped (P0)	Marker data missing for top generation and 50% of middle generation (M2)	82.2	57.7	35.2
4	Affected individuals haplotyped (P1)	No marker data missing (M0)	97.1	89.4	74.2
5	Affected individuals haplotyped (P1)	Marker data missing for top generation (M1)	96.7	87.1	71.2
6	Affected individuals haplotyped (P1)	Marker data missing for top generation and 50% of middle generation (M2)	92.0	75.4	55.1
7	All members haplotyped (P2)	No marker data missing (M0)	97.5	90.0	75.2
8	All members haplotyped (P2)	Marker data missing for top generation (M1)	97.3	89.2	73.5
9	All members haplotyped (P2)	Marker data missing for top generation and 50% of middle generation (M2)	94.5	79.4	62.4

of a high-density map of SNP markers compared with a less dense[25–29] set of STR markers. However, most laboratories today are using a twofold more dense SNP map (1 SNP every 0.5 cM) than the one simulated in this study; therefore, it is inappropriate to use these data to compare the power of using STR versus SNP markers in a GWS.

It should be noted that our analyses with VITESSE version 2 assume not only a molecular haplotype but also a parental source for each haplotype. Although there are a priori two possible parental sources for each molecular haplotype, parental phase can be established with a single informative transmission, because the parental source of a single allele identifies the parental source of all alleles on the haplotype. With a sufficiently dense set of polymorphic markers, we believe it is reasonable to assume that parental phase can be established in an individual. Using this knowledge of phase and a sufficiently dense set of polymorphic markers to identify regions shared identical by descent (IBD) from the same parent, we can assume that the parental source is also known in the individual's siblings. If the parents are not genotyped, then parental source cannot be determined, but it can be arbitrarily assumed, provided both parents have the same phenotype. However, in practice, if untyped parents have different phenotypes or if two siblings share no alleles IBD along the chromosome, then it may be necessary to consider analyses for both possible parental phases.

Complex diseases account for most of the public health burden. Characterization of the genetic factors involved in disease etiology has proven beneficial in the diagnosis, prognosis, and treatment of disease. However, efforts to understand molecular pathogeneses of these diseases have been limited. This study provides further evidence of the increased value of molecular haplotyping in the context of linkage studies. However, it must be balanced against the considerable cost and effort of molecular haplotyping methodologies. At today's costs, these methods would be economically feasible only to follow up a few regions of suggestive linkage when collection of additional families is difficult. However, these positive results should encourage novel molecular techniques, which could facilitate the use of molecular haplotyping information. In particular, maximizing power given an existing family set can be particularly important in late-onset, often-fatal traits such as pancreatic or lung cancer, for which informative families are difficult to collect.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. We thank Drs. Gloria Petersen, Bob Jenkins, and Ed Highsmith for their valuable input regarding the cost of conversion, which will be offered as a service at the Mayo Clinic. We also thank Alejandro Schaffer for early discussions as well as sharing a modified version of mlink (phaselink).

References

1. Puffenberger EG, Kauffman ER, Bolk S, Matise TC, Washington SS, Angrist M, Weissenbach J, Garver KL, Mascari M, Ladda R, Siaugenhaupt S, Chakravarti A (1994) Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum Mol Genet 3: 1217–1225
2. Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region β 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97:10483–10488
3. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603
4. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228
5. Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959
6. Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequen-

cies: an example from the CD4 locus. Am J Hum Genet 67: 518–522
7. Kirk KM, Cardon LR (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur J Hum Genet 10:616–622
8. Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24: 4841–4843
9. Antonellis A, Rogus JJ, Canani LH, Makita Y, Pezzolesi MG, Nam M, Ng D, Moczulski D, Warram JH, Krolewski AS (2002) A method for developing high-density SNP maps and its application at the type 1 angiotensin II receptor (AGTR1) locus. Genomics 79:326–332
10. Yu CE, Devlin B, Galloway N, Loomis E, Schellenberg GD (2004) ADLAPH: a molecular haplotyping method based on allele-discriminating long-range PCR. Genomics 84:600–612
11. Papadopoulos N, Leach FS, Kinzler KW, Vogelstein B (1995) Monoallelic mutation analysis (MAMA) for identifying germline mutations. Nat Genet 11:99–102
12. Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56:303–320
13. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361–364
14. Schaid DJ (2002) Relative efficiency of ambiguous vs directly measured haplotype frequencies. Genet Epidemiol 23:426–443
15. Thomas S, Porteous D, Visscher PM (2004) Power of direct vs indirect haplotyping in association studies. Genet Epidemiol 26:116–124
16. Wilson AF, Bailey-Wilson JE, Pugh EW, Sorant AJM (1996) The Genometric Analysis Simulation Program (GASP): a software tool for testing and investigating methods in statistical genetics. Am J Hum Genet 59:A193
17. O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408
18. O'Connell JR (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 51:226–240
19. Ott J (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore
20. Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129–1137
21. Kamio K, Matsushita I, Tanaka G, Ohashi J, Hijikata M, Nakata K, Tokunaga K, Azuma A, Kudoh S, Keicho N (2004) Direct determination of MUC5B promoter haplotypes based on the method of single-strand conformation polymorphism and their statistical estimation. Genomics 84:613–622
22. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
23. Weeks DE, Sobel E, O'Connell JR, Lange K (1995) Computer programs for multilocus haplotyping of general pedigrees. Am J Hum Genet 56:1506–1507
24. Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am J Hum Genet 71:992–995
25. John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W, Vasavda N, Mills T, Barton A, Hinks A, Eyre S, Jones KW, Ollier W, Silman A, Gibson N, Worthington J, Kennedy GC (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. Am J Hum Genet 75:54–64
26. Middleton FA, Pato MT, Gentile KL, Morley CP, Zhao X, Eisener AF, Brown A, Petryshen TL, Kirby AN, Medeiros H, Carvalho C, Macedo A, Dourado A, Coelho I, Valente J, Soares MJ, Ferreira CP, Lei M, Azevedo MH, Kennedy JL, Daly MJ, Sklar P, Pato CN (2004) Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide–polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am J Hum Genet 74:886–897
27. Sawcer SJ, Maranian M, Singlehurst S, Yeo T, Compston A, Daly MJ, De Jager PL, Gabriel S, Hafler DA, Ivinson AJ, Lander ES, Rioux JD, Walsh E, Gregory SG, Schmidt S, Pericak-Vance MA, Barcellos L, Hauser SL, Oksenberg JR, Kenealy SJ, Haines JL (2004) Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. Hum Mol Genet 13:1943–1949
28. Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hilker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML, Thibodeau SN (2004) Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer–susceptibility loci. Am J Hum Genet 75:948–965
29. Klein AP, Tsai YY, Duggal P, Gillanders EM, Barnhart M, Mathias RA, Dusenberry IP, Turiff A, Chines PS, Goldstein J, Wojciechowski R, Hening W, Pugh EW, Bailey-Wilson JE (2005) Investigation of altering single-nucleotide polymorphism density on the power to detect trait loci and frequency of false positive in nonparametric linkage analyses of qualitative traits. BMC Genet Suppl 6:S20