Published in final edited form as:

Genet Epidemiol. 2008 January ; 32(1): 41–51.

Examining the effect of linkage disequilibrium between markers on the Type I error rate and power of nonparametric multipoint linkage analysis of two-generation and multigenerational pedigrees in the presence of missing genotype data

Yoonhee Kim^{1,2}, Priya Duggal², Elizabeth M. Gillanders², Ho Kim^{1,*}, and Joan E. Bailey-Wilson^{2,*}

¹ Department of Biostatistics and Epidemiology, School of Public Health, Seoul National University, Seoul, Republic of Korea

² Inherited Disease Research Branch, NHGRI/NIH, Baltimore, MD, U.S.A

Abstract

Since most multipoint linkage analysis programs currently assume linkage equilibrium (LE) between markers when inferring parental haplotypes, ignoring linkage disequilibrium (LD) may inflate the Type I error rate. We investigated the effect of LD on the Type I error rate and power of nonparametric multipoint linkage analysis of two-generation and multigenerational multiplex families. Using genome wide single nucleotide polymorphism (SNP) data from the Collaborative Study of the Genetics of Alcoholism (COGA), we modified the original dataset into 30 total data sets in order to consider 6 different patterns of missing data for 5 different levels of SNP density. To assess power, we designed simulated traits based on existing marker genotypes. For the Type I error rate, we simulated 1,000 qualitative traits from random distributions, unlinked to any of the marker data. Overall, the different levels of SNP density examined here had only small effects on power (except sibpair data). Missing data had a substantial effect on power, with more completely genotyped pedigrees yielding the highest power (except sibpair data). Most of the missing data patterns did not cause large increases in the Type I error rate if the SNP markers were more than 0.3 cM apart. However, in a dense 0.25 cM map, removing genotypes on founders and/or founders and parents in the middle generation caused substantial inflation of the Type I error rate, which corresponded to the increasing proportion of persons with missing data. Results also showed that long high-LD blocks have severe effects on Type I error rates.

Keywords

SNPs; Type I error rate; False Positives; Linkage Disequilibrium; Pedigree Structure

Email addresses of authors YK : nina78@snu.ac.kr PD : pduggal@mail.nih.gov EMG : lgilland@mail.nih.gov *HK: hokim@snu.ac.kr

*JEBW : jebw@mail.nih.gov

*co-corresponding authors

HK School of Public Health, Seoul National University #28 Yeunkun dong Chongno gu Seoul 110-799, Republic of Korea TEL (+822) – 740 – 8874, FAX (+822) – 745 – 9104 hokim@snu.ac.kr

JEBW IDRB/NHGRI/NIH 333 Cassell Drive, Suite 1200 Baltimore, MD 21224, U.S.A TEL (410) – 550 – 7509, FAX (410) – 550 – 7513 jebw@mail.nih.gov

Introduction

High-density, genome wide, SNP panels have become major and practical resources for linkage analyses, because of the technical advantages of SNPs compared with microsatellite markers and the reduced cost of genotyping [Evans and Cardon 2004; Goode and Jarvik 2005; Schaid, et al. 2004]. Although each SNP is less informative than a microsatellite marker, the higher density of SNP markers when used in multipoint linkage analysis can compensate for this individual weakness in informativeness [Browning, et al. 2004; Klein, et al. 2005; Murray, et al. 2004; Schaid, et al. 2004]. For example, a typical SNP mean density in a linkage panel may be 1 SNP every 0.6 cM, as compared to microsatellite marker maps which typically use 1 marker about every 10 cM [Evans and Cardon 2004].

In general, increased density of markers predicts stronger intermarker linkage disequilibrium (LD), although these are not perfectly correlated. Problems using SNP markers for linkage analysis have arisen when dense panels of SNPs are used, because there is often LD between these markers and most analysis programs currently assume linkage equilibrium. Previous studies [Boyles, et al. 2005; Goode, et al. 2005; Goode and Jarvik 2005; Huang, et al. 2004; Huang, et al. 2005; Schaid, et al. 2004] have shown that high-density SNPs in LD could increase the Type I error rate in multipoint linkage analysis. To calculate the LOD score in multipoint linkage analysis, the IBD (identity by descent) status of markers has to be estimated based on allele frequencies using possible relative pairs. The LD between markers could interfere with inferring parental haplotype frequencies for estimating IBD sharing. Therefore, increased estimates of IBD sharing due to misspecified haplotype frequencies lead to increased false positive rates for linkage [Huang, et al. 2004]. If parental genotypes are unavailable, as is the case in most late onset diseases, this problem will be worse. Several studies [Boyles, et al. 2005; Huang, et al. 2004; Levinson and Holmans 2005] investigated the changes of the Type I error rate in nuclear families with parental genotypes missing, while considering various degrees of LD using r^2 or D' , and various minor allele frequencies at the SNP marker loci. These studies showed that missing parental genotypes increased the Type I error rate.

A number of options to control the Type I error rate when analyzing data with missing parental genotypes in the presence of LD have been suggested. Boyles et al. and Huang et al. [2005; 2004;2005] showed that adding either flanking markers in equilibrium or additional unaffected siblings may decrease but not eliminate the inflation of the Type I error rate. Bacanu [2005] proposes partitioning the markers and performing multipoint linkage analysis using subsets of adjacent SNP markers with weak LD. Additionally, there are software programs, SNPLINK [Webb, et al. 2005] and MERLIN v1.0.0 [Abecasis, et al. 2002], that allow for this problem. SNPLINK removes markers in LD automatically prior to nonparametric or parametric linkage analysis, and MERLIN has an option of haplotype sampling to produce correct results in the presence of LD. More recently Xing et al. [2006] advocated using discordant sibpairs to manage multipoint nonparametric linkage analysis of affected sib pair studies with intermarker LD.

To evaluate the effect of both intermarker LD and missing genotype data, previous studies have considered various degrees of LD, minor allele frequencies, and parental missing genotype data. However, these studies have examined only small regions of the genome with a few markers, most often using simulated genotype data for nuclear families [Boyles, et al. 2005; Huang, et al. 2004; Huang, et al. 2005; Levinson and Holmans 2005]. In human genetics there exist unpredictable differences among individuals, families, and risk factors for disease. Evaluating the effect of intermarker LD on a multipoint linkage analysis based on simple pedigree structures using simulated genotype data with restrictive assumptions could lack generality. In particular, genome wide LD patterns, even with extremely delicate modeling of parameters such as LD and allele frequency using present simulation programs, might not mimic realistic LD patterns in a linkage panel of SNPs. In this paper, we investigated genome

wide Type I error rate and power in multipoint nonparametric linkage analysis of multigenerational pedigrees across five levels of SNP density (1 SNP every 0.25, 0.3, 0.6, 1, or 2 cM) using actual genotype data which present real, moderate LD patterns in human subjects. Using family data from the Collaborative Study of the Genetics of Alcoholism (COGA) we modified the original multiplex families into 5 additional data sets with different patterns of missing genotype data and different pedigree structures, to assess the effects of SNP density on power and the Type I error rate. Our goal is to identify the optimal density of SNP markers for a linkage study that balances the increased informativeness of dense SNPs with the potential increase in the Type I error rate and to determine whether the effects of LD on the Type I error rate differ according to proportion of missing parental genotypes and/or pedigree structure.

Materials and Methods

Pedigree Data

We used multiplex family data from the Collaborative Study of the Genetics of Alcoholism (COGA) including nuclear and multigenerational families originally ascertained from alcohol-dependent probands. The initial research purpose and data collection procedure are well documented elsewhere [Reich, et al. 1998]. We did not use any of the phenotype data from these families, but instead constructed our own simulated traits as described below. The existing SNP genotype data were used in our analyses. To avoid allele frequency differences among families due to ethnic diversity, we limited our study to white/non-Hispanic families (102 families) from the 143 families genotyped and provided by Genetic Analysis Workshop 14 (GAW14) [Edenberg, et al. 2005]. These 102 families (40 two-generational families and 62 multigenerational families) consisting of 938 individuals had 139 missing founder genotypes and 30 missing non-founder genotypes. Using this original (denoted ORIGINAL) data, we created 5 additional data sets: the first 3 data sets are the original multigenerational pedigree data (1) removing one parental genotype at random in each mating in the top and middle generations (ONE, note: if both parents in a mating were already ungenotyped, then they remained that way for the “ONE” dataset), (2) removing all founders' genotypes in the top generation (TOP), (3) removing both parental genotypes in each mating in the top and middle generations (TWO). In the fourth dataset (4) we reorganized the data into nuclear families having two or more siblings without both parental genotypes (NUCLEAR). The original 102 multiplex families yielded 205 nuclear families and we deleted all parental genotypes from these families. Finally, the fifth dataset (5) reorganized the nuclear family data into single sib pairs without both parental genotypes (SIBPAIR). Two nuclear families having only a single child were deleted because this type of family structure provides no information for linkage analysis. To make the SIBPAIR data set, we determined all possible sibling pairs in the nuclear families and made each sibling pair a unique family, resulting in 948 sib-pair families. The basic pedigree information for the 6 data sets is summarized in Table I. In addition, supplementary Table III and supplementary Figure 1 illustrate how these data sets were created.

Genetic markers

To obtain SNP maps with diverse scales of densities from very dense (an average of 1 SNP every 0.25 cM) to less dense (an average of 1 SNP every 2 cM), we merged 11,560 SNP markers from the Affymetrix 10K assay with 4,600 SNP markers from Illumina, which mapped to unique physical positions based on the Single Nucleotide Polymorphism data base (dbSNP) of the National Center for Biotechnology Information (NCBI build 34), as previously described [Duggal, et al. 2005; Klein, et al. 2005]. This Affymetrix linkage analysis panel has a mean genetic distance of 0.36 cM (<http://www.affymetrix.com/products/arrays/specific/10k2.affx>), and the Illumina panel has a mean genetic distance of 0.64 cM.

(http://www.illumina.com/products/prod_snp.ilmn). The integrated high-density map, (denoted “0.25 cM” map), consisted of 15,019 unique SNPs with a mean intermarker distance of 189 kb. This original 0.25 cM map had regions of strong pairwise intermarker linkage disequilibrium (LD) showing D' of 0.7 or greater. $D' > 0.7$ was observed for 21% of all SNP pairs that were within 500kb of each other (44,026 pairs) [Beckmann, et al. 2005]. After the integration of the 2 different panels, we generated 4 additional SNP density levels by selecting a subset of SNPs, explicitly, one marker every 0.3, 0.6, 1, or 2 cM, while choosing markers with major allele frequency closest to 0.5 among multiple markers within $\pm 10\%$ of the desired distance. Per the LD maps of each SNP density (data not shown), LD blocks are observed most frequently along the chromosomes for the 0.25 cM density map. The numbers and sizes of LD blocks decrease as SNP density decreases. However, this pattern is not numerically linear to SNP densities since SNP density represents the average density of the markers. For example, there is a dramatic reduction in intermarker LD between the 0.25 cM and the 0.3 cM maps. In the 0.3 cM map (5,405 SNPs), we have 4,124 pairwise comparisons within 500kb of each other and 148 (3.6%) had a $D' > 0.7$.

Qualitative Traits

To assess the effect of LD on power of multipoint linkage analysis, we examined the Kong & Cox LOD score and the p-value of the NPL score (S_{all}) at simulated “causative” loci by creating traits generated from the genotypes of known SNP loci. We created 20 dominant qualitative traits using the following markers as the known causative locus: rs1796969 (chr1), tsc0583756 (chr2), rs269384 (chr3), rs1051447 (chr4), rs34999 (chr5), rs159988 (chr6), rs42611 (chr7), rs765262 (chr8), rs363717 (chr9), tsc0040918 (chr10), tsc0912310 (chr11), tsc0591807 (chr12), tsc0054081 (chr13), tsc0914256 (chr14), rs488756 (chr15), tsc0513689 (chr16), rs757288 (chr17), rs1941207 (chr18), rs1603 (chr19), and tsc0046507 (chr20). The traits were named by placing a prefix “D” in front of each marker name to avoid confusion with the marker names. These loci were picked to create dominant disease traits because their minor allele frequency (MAF) was between 10% and 20%, which resulted in each pedigree being adequately well-balanced in terms of affected and unaffected family members. For each created trait, the allele with the lowest allele frequency at the SNP locus was treated as the “risk” allele and the common allele was treated as the “normal” allele. People who had 1 or 2 copies of the “risk” allele at the chosen SNP locus were coded as affected with complete penetrance (100%). People who had 2 copies of the “normal” allele were coded as unaffected at the generated trait. Individuals not genotyped in the original data set were coded as unknown. This process was repeated for each of the SNP loci listed above, creating 20 generated dominant traits that are “caused by” the generating SNP and thus linked to the region of the causative SNP.

For the Type I error rate, we simulated 1,000 qualitative traits from random distributions, which were unlinked to all of the marker loci. In other words, all of these traits are randomly simulated, so that there cannot be relationships between the trait values between relatives or between the trait and the SNP markers. One thousand independent traits were simulated with 40% (333 traits), 60% (333 traits), or 80% (334 traits) prevalence rate among 938 individuals, which resulted in each pedigree being adequately well-balanced in terms of affected and unaffected family members. For each trait, the trait status was simulated for each person by choosing a random number from a Uniform (0,1) distribution using the random number generator in the statistical software package R v2.3.1, and designating the person as affected if their random number was less than each fixed prevalence rate, and unaffected otherwise.

Multipoint nonparametric linkage (NPL) analysis

Thirty sets of multipoint nonparametric linkage analyses performed on 6 data sets with different missing data patterns/pedigrees for 5 different levels of SNP density were conducted on the 1,020 simulated traits described above, using Merlin v0.10.2 [Abecasis, et al. 2002] with allele

frequencies estimated from all founders. If founders were not available, Merlin automatically predicted allele frequencies using all individuals. The information content of the genotypes was estimated by use of the entropy information described by Kruglyak et al. [1996] using Merlin, and we averaged the information content across markers (Results in Supplementary Table I). Kong and Cox [1997] allele-sharing LOD scores and the p-values for the NPL [Whittemore and Halpern 2003] using the score function of S_{all} were used for the model free qualitative linkage analysis. Using the maximum Kong and Cox LOD score and minimum NPL p-values within 20 cM of the true trait locus, we compared power to detect linkage of the 20 simulated “causative” loci in each data set for each marker map density.

For the Type I error rate, we counted the number of peaks having p-values less than 4 standard thresholds (0.05, 0.01, 0.001, and 0.0001) and 2 Lander-Kruglyak [Lander and Kruglyak 1995] genome wide significance levels (0.0017, and 0.000049) over all chromosomes for each trait, and then summed over all 1,000 traits. To avoid counting the same peak (stretching over several flanking markers) multiple times, we required that a p-value greater than or equal to 0.2 must occur between each independent false positive region. For example, if a p-value less than 0.000049 occurred on chromosome 1 at 5 cM from the telomere, with each location for the next 2 cM also showing p-values less than or equal to 0.000049, we would count all these p-values as indicative of only 1 false positive linkage peak. If, over the next 10 cM the p-value rose to 0.2 or greater and then another p-value of less than or equal to 0.000049 was observed, we would count this new location as yielding an independent false positive linkage signal at the $p=0.000049$ significance level.

Results

Power comparison by LOD scores and p-values observed in the multipoint linkage analysis

Table II displays the LOD scores and p-values observed in the linkage analysis of the 20 generated traits to the region containing the “causative” locus across 6 data sets using the lowest density map of 2 cM SNP markers and the highest density map of 0.25 cM SNP markers. For 19/20 traits (except Dtsc0583756 trait, LOD=1.68, p-value=0.003) LOD scores larger than 3 were observed in the region of their “causative” locus in both the 2 cM density SNP marker set and the 0.25 cM density SNP marker set. Overall, LOD scores increased when using the 0.25 cM density SNP marker set compared to the 2 cM density SNP marker set, although the effect of increasing SNP density does not show a consistent effect on the size of the LOD score across all data structures (Figure 1). The largest proportional increase was seen in the SIBPAIR data set (Table II, Supplementary Table II).

In general, the effect of missing data was greater than the effect of SNP density on power. The 19 traits with good power in the ORIGINAL data also show reasonable power (LOD scores usually greater than 3) in the ONE, TOP, and TWO data sets. However, in the NUCLEAR and SIBPAIR pedigree structure data sets, as expected the number of traits which have LOD scores greater than 3 was decreased compared to the multigenerational data sets (Table II, Supplementary Table II). Additionally, for 15/20 traits in all 5 densities of SNP markers as shown in figure 1, the power was higher in the multigenerational pedigree structures compared to the NUCLEAR or SIBPAIR data sets.

Type I error rate

Table III displays the number of false positive signals below standard p-values (0.05, 0.01, 0.001, and 0.0001) and genome wide suggestive and significant p-values (0.0017, and 0.000049) of the NPL statistics for each missing data pattern/pedigree and SNP density level for 1,000 simulated traits. It is not surprising that false positive signals decreased steeply as the significance levels decreased. Using p-value thresholds at the commonly recommended

0.0001 and 0.000049 levels helps to control the genome wide error rate due to the multiple testing problem in multipoint linkage analysis. However, even when examining the results limited to the more stringent significance thresholds (p-values less than or equal to 0.0001, and 0.000049), it is clear that an increasing density of SNP markers increases the Type I error rate. Appropriate or conservative Type I error rates were observed for four of the data sets (ORIGINAL, ONE, TOP, and TWO) when using the 0.3 cM, 0.6 cM, 1 cM, and 2 cM densities, with a clear reduction of the Type I error rate when using the 2 cM density map compared to the 1 cM density map. However, all data structures showed some inflation of genome wide Type I error rate when the 0.25 cM map was used (expected 5% Type I error rate), with the ORIGINAL and ONE data sets showing only mild inflation (6% or 7.7% of the traits showed Type I errors at the genome wide significance level, $p=0.000049$). The TOP, TWO, and NUCLEAR datasets showed moderate inflation (about 20% of the traits showed Type I errors at the genome wide significance level) (Figure 2). However, remarkable inflation of the Type I error rate was found in the SIBPAIR data set; an average of 2 significant peaks per genome wide scan (i.e., 2,015 false positives/1,000 genome scans=2.015) were observed whereas only 1 significant peak in 20 genome scans would be expected. One intriguing observation in the SIBPAIR data set was that the 2 cM SNP marker map also showed a larger number of false positive peaks (slightly over 20%) than the three maps of intermediate density.

Closer examination of the locations of the Type I errors in the different data sets revealed that the majority of errors for the 0.25 cM map occurred on chromosomes 6, 7, 12, 19, 20 and 21 (Supplementary Table IV shows the chromosome specific frequency of the Type I errors for the 0.25 cM map). The majority of the Type I errors observed for this map in the TOP, TWO and NUCLEAR data sets was due to errors on Chromosome 21 (Supplementary Table IV). However, even when the errors on Chromosome 21 are ignored, these data sets still have inflated Type I error rates when using the 0.25 cM map. Chromosomes 6, 12, 20 and 21 no longer show excessive Type I errors once the map density decreases to 0.3cM (Supplementary Tables V and VI). Chromosome 21 has a very large high-LD block in the 0.25cM map which no longer exists in the 0.3 cM or less dense maps (Supplementary Figure 4). Chromosomes 6, 7, 12, 19, and 20 have a large number of short high-LD blocks in the 0.25 cM map that are markedly reduced in number in the less dense maps (Supplementary Figures 1-3 show examples of typical LD structure in these data). However, the increased Type I errors continue to be observed on chromosome 7 even in the least dense map for the ORIGINAL data set (Supplementary Table V) and on chromosome 19 for the 0.3 and 0.6 cM maps in the NUCLEAR data set (Supplementary Table VI).

Discussion

Since standard genome wide SNP linkage panels may retain substantial LD between markers, and high-density marker maps are often genotyped during fine mapping or in association studies, and these same genotypes may also be used for linkage analyses, concern about inflated false positive linkage rates induced by LD has become an important issue in multipoint linkage analysis. We investigated frequencies of false positive peaks and evaluated the power in model-free multipoint linkage analysis in 6 data structures with varying amounts of missing genotype patterns/pedigrees, using 5 densities of genome wide SNP marker maps. The strengths of our study were 1) that we used real genome wide genotype data containing moderate genome wide LD patterns and multigenerational pedigree structures that reflect realistic data structures commonly used in general linkage analysis and 2) that we were able to compare different patterns of missing genotype data/pedigree structures and different map densities within the same data.

There was no simple relationship between power and the data structures examined here, or between power and the density of SNP markers for all simulated traits. However, in general,

for the majority of traits the more completely genotyped pedigrees were more powerful. In addition, the more dense maps had somewhat higher power than the less dense maps. This could be explained by the fact that the trait generating locus was always included in the most dense map. The most striking increases in power with increasing SNP density were in the SIBPAIR data. This could be the result of increase in information content (0.29-0.49) between the maps (2 cM-0.25 cM) for the SIBPAIR data set. These striking increases in power could also reflect, in part, inflation of test statistics due to increasing intermarker LD, since this inflation appears to be most severe in the SIBPAIR data set. In this study, we cannot differentiate between these possibilities.

The Type I error rate shows more apparent impact from alteration of the densities of the SNP markers across different levels of missing genotype data and different pedigree structures. Our results, assuming that degree of LD might be higher if SNP markers are denser, confirm and extend the previous studies on effects of intermarker LD in linkage analyses that assume linkage equilibrium. In general, we found that a very dense map (0.25 cM) caused inflation of the Type I error rate in all data structures, with the severity increasing as the amount of missing genotype data increased and as the pedigree structure changed from multigenerational to nuclear family to sibpair. Huang et al. and Evans and Cardon have suggested that adding more unaffected sibs to datasets consisting of independent affected sibpairs with missing parental genotypes would be a solution to control the Type I error rate [Evans and Cardon 2004; Huang, et al. 2004]. Indeed, our nuclear family framework performed better than the single sib pair data without parental genotypes, in that the Type I error rate was not as badly inflated in the larger nuclear families for the densest 2 maps and was not inflated at all for the 0.6 cM, 1 cM and 2 cM maps. In addition, larger, multigenerational pedigree structures can play a role to minimize the combined effects of intermarker LD and missing genotypes within each pedigree. In the multigenerational pedigrees examined here, the Type I error rate was not inflated even when there were more missing genotypes in each pedigree when using the 0.3 cM, 0.6 cM, 1 cM, and 2 cM marker maps. In contrast, the sib pair data without parental genotypes showed inflation of the Type I error rate when using any density of marker map tested here. The extreme inflation in the SIBPAIR data set could partially be due to violation of the assumption of independence of the sib pairs. Interestingly, in the SIBPAIR data set, the 0.25 cM map showed the strongest inflation but the 2 cM map showed a higher inflation of the Type I error rate than did the intermediate density maps.

We traced back the locations of the Type I errors that had p-values of 0.000049 or less in the 0.25cM map, and observed many false positive peaks in regions where there were multiple blocks of markers in high intermarker LD. As the blocks of high intermarker LD disappeared in the less dense maps, the Type I errors generally also decreased. However, two regions, on chromosomes 7 and 19, were unusual (Supplementary Tables V-VIII). On chromosome 7 in the ORIGINAL data (Supplementary Table VIII), Type I errors continued to be high across all map densities, even when virtually no intermarker LD was observed. This could be due to a limitation of our study: we used a single set of genotype data from the COGA /GAW14 study for all 1,000 simulated traits. The COGA data were ascertained based on alcohol-dependence and linkage to alcohol-dependence related traits has been reported at chromosomes 7 and 11 [Daw, et al. 2005]. We suspect the high number of the Type I errors on chromosome 7 in our study may be due to chance similarities in affection status between the original traits on which these families were ascertained and the affection statuses for the simulated traits in a few of the simulated replicates. Thus the increased Type I errors in this region could reflect allele sharing due to linkage of the original traits in this data set. On chromosome 19, there were increased Type I errors in the 0.3 and 0.6 cM maps for the NUCLEAR data set but they appeared to occur in regions that did not have many markers in high LD with each other. The cause of these Type I errors is not easily explained, except that Type I errors are a stochastic process and thus this less striking pattern may not be meaningful.

In conclusion, we recommend the following be considered in nonparametric multipoint linkage analysis using genome wide SNP markers. Very dense SNP marker maps do not have clear advantages for increasing power, once a density of 1 SNP every 1 or 2 cM is achieved. Therefore, we recommend focusing on minimizing the Type I error rate rather than considering only power when determining the density of SNP markers. It is common practice to use 0.0001 or 0.000049 as thresholds of p-values for declaring significance, in order to control the family-wise Type I error rate. Most of the incompletely genotyped data structures had similar levels of Type I error rate at the $p=0.000049$ level if the SNP markers were spaced more than 0.3 cM apart, except the sib pair data (which incorrectly assumed independence among the sibling pairs). However, the very dense map (0.25 cM) that exhibited moderate levels of intermarker LD caused severe inflation of the Type I error rate when both parental genotypes were missing, in the top generation (TOP) or in both the top and middle generations in the multigenerational pedigree data sets (TWO), with smaller inflation being observed when only one parental genotype was missing in most matings (ONE). Addition of either siblings or relative genotypes appears to be helpful in controlling the Type I error rate at all of the SNP density levels examined here, but it is clear that even in well genotyped multigenerational pedigrees, moderate levels of intermarker LD can cause moderate increases in the genome wide Type I error rate. At present, methods to take LD into account in linkage studies are only practical for analyzing small regions of the genome at one time. Current genome wide SNP linkage panels have an average density of 0.36 cM and 0.64 cM (Affymetrix and Illumina, respectively), which may result in increased false positive linkage peaks in nuclear family or sibpair data. While the effects on Type I error rate are less striking in multigenerational pedigrees, it is clear that strong intermarker LD can still cause inflated false positive rates particularly when many parental genotypes are missing. Further studies are needed to elucidate how much intermarker LD is sufficient to cause false positive peaks. Our results suggest that large numbers of short high-LD blocks are adequate to increase Type I error in multigenerational and nuclear families with missing parental data. We also showed that long high-LD blocks have severe effects on the Type I error. Therefore, we recommend that researchers always check for intermarker LD in their SNP marker panels and either remove markers prior to analyses until levels of intermarker LD are low or that they carefully reanalyze any significant linkage peaks using methods that account for intermarker LD. Identification of disease susceptibility genes is a lengthy and expensive venture, and all measures to control for incorrect findings should be taken prior to the publication of any linkage results using dense SNP maps.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was partly supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF 2005-213-C00007) and in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Data were provided by the Genetic Analysis Workshop 14 (supported by NIH grant GM31575) and by the Collaborative Study on the Genetics of Alcoholism (COGA; supported by the NIH Grant U10AA08403 from the National Institute on Alcohol Abuse and Alcoholism and the National Institute on Drug Abuse). CIDR Genotyping services for these data for GAW14 were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number N01-HG-65403.

References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30(1):97–101. [PubMed: 11731797]
- Bacanu SA. Multipoint linkage analysis for a very dense set of markers. *Genet Epidemiol* 2005;29(3): 195–203. [PubMed: 16094613]

- Beckmann L, Ziegler A, Duggal P, Bailey-Wilson JE. Haplotypes and haplotype-tagging single-nucleotide polymorphism: presentation Group 8 of Genetic Analysis Workshop 14. *Genet Epidemiol* 2005;29(Suppl 1):S59–71. [PubMed: 16342175]
- Boyles AL, Scott WK, Martin ER, Schmidt S, Li YJ, Ashley-Koch A, Bass MP, Schmidt M, Pericak-Vance MA, Speer MC. Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered* 2005;59(4):220–7. [PubMed: 16093727]others
- Browning BL, Brashear DL, Butler AA, Cyr DD, Harris EC, Nelsen AJ, Yarnall DP, Ehm MG, Wagner MJ. Linkage analysis using single nucleotide polymorphisms. *Hum Hered* 2004;57(4):220–7. [PubMed: 15583428]
- Daw EW, Doan BQ, Elston RC. Linkage mapping methods applied to the COGA data set: presentation Group 4 of Genetic Analysis Workshop 14. *Genet Epidemiol* 2005;29(Suppl 1):S29–34. [PubMed: 16342182]
- Duggal P, Gillanders EM, Mathias RA, Ibay GP, Klein AP, Baffoe-Bonnie AB, Ou L, Dusenberry IP, Tsai YY, Chines PS. Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium. *BMC Genet* 2005;6(Suppl 1):S73. [PubMed: 16451687]others
- Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R, Doheny KF, Hansen M, Hinrichs T, Jones K. Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genet* 2005;6(Suppl 1):S2. [PubMed: 16451628]others
- Evans DM, Cardon LR. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet* 2004;75(4):687–92. [PubMed: 15311375]
- Goode EL, Badzioch MD, Jarvik GP. Bias of allele-sharing linkage statistics in the presence of intermarker linkage disequilibrium. *BMC Genet* 2005;6(Suppl 1):S82. [PubMed: 16451697]
- Goode EL, Jarvik GP. Assessment and implications of linkage disequilibrium in genome-wide single-nucleotide polymorphism and microsatellite panels. *Genet Epidemiol* 2005;29(Suppl 1):S72–6. [PubMed: 16342185]
- Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* 2004;75(6):1106–12. [PubMed: 15492927]
- Huang Q, Shete S, Swartz M, Amos CI. Examining the effect of linkage disequilibrium on multipoint linkage analysis. *BMC Genet* 2005;6(Suppl 1):S83. [PubMed: 16451698]
- Klein AP, Tsai YY, Duggal P, Gillanders EM, Barnhart M, Mathias RA, Dusenberry IP, Turiff A, Chines PS, Goldstein J. 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* 2005;6(Suppl 1):S20. [PubMed: 16451629]others
- Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997;61(5):1179–88. [PubMed: 9345087]
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58(6):1347–63. [PubMed: 8651312]
- Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995;11(3):241–7. [PubMed: 7581446]
- Levinson DF, Holmans P. The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees. *BMC Genet* 2005;6(Suppl 1):S6. [PubMed: 16451672]
- Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan JB, Chee MS. A highly informative SNP linkage panel for human genetic studies. *Nat Methods* 2004;1(2):113–7. [PubMed: 15782173]others
- Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, Bucholz K. Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet* 1998;81(3):207–15. [PubMed: 9603606]others
- Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hilker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility Loci. *Am J Hum Genet* 2004;75(6):948–65. [PubMed: 15514889]others

- Webb EL, Sellick GS, Houlston RS. SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. *Bioinformatics* 2005;21(13):3060–1. [PubMed: 15840706]
- Whittemore AS, Halpern J. Genetic association tests for family data with missing parental genotypes: a comparison. *Genet Epidemiol* 2003;25(1):80–91. [PubMed: 12813729]
- Xing C, Sinha R, Xing G, Lu Q, Elston RC. The affected-/discordant-sib-pair design can guarantee validity of multipoint model-free linkage analysis of incomplete pedigrees when there is marker-marker disequilibrium. *Am J Hum Genet* 2006;79(2):396–401. [PubMed: 16826532]

FIGURE 1. LOD scores of 20 traits depending on SNP densities, as proportion of missing parental genotypes and pedigree structures change

X axis: density of SNP markers (*: “causative” marker is included in this marker map), Y axis: Kong & Cox LOD score, Different data structures; Black solid line: ORIGINAL - initial data set in multigenerational family structure, Red long & short dotted line: ONE- one parental genotype was removed at each marker locus in top and middle generations, Green dotted line: TOP - both founder genotypes were removed at each marker locus in top generation, Blue dotted & long line: TWO- all parental genotypes were removed at each marker locus in top and middle generations, Sky blue broken line: NUCLEAR- nuclear families without two parental genotypes at all marker loci, and Pink short & broken line: SIBPAIR – single sib pairs without parental genotypes at all marker loci in nuclear family framework.

FIGURE 2. Number of Type I errors (p-value <0.000049) in each data set as SNP densities are varied

X axis: density of SNP markers, Y axis: Number of Type I errors for each data set, Different data structures; Solid line: ORIGINAL - initial data set in multigenerational family structure, Long & short dotted line: ONE- one parental genotype was removed at each marker locus in top and middle generations, Dotted line: TOP - both founder genotypes were removed at each marker locus in top generation, Dotted & long line: TWO- all parental genotypes were removed at each marker locus in top and middle generations, Broken line: NUCLEAR- nuclear families without two parental genotypes at all marker loci, and Short & broken line: SIBPAIR – single sib pairs without parental genotypes at all marker loci in nuclear family framework, *SIBPAIR line is truncated at 0.3 cM, because the 2,000 Type I errors observed for the 0.25 cM marker map exceed the margin of the figure.

TABLE I

Pedigree information of six data sets

	ORIGINAL [*] ONE TOP Δ TWO	NUCLEAR [†]	SIBPAIR [‡]
Pedigree number	102	205	948
Mean size	9.22+/-2.71	5.08+/-1.68	4
Sibships	205	205	948
Mean size	3.08+/- 1.68	3.08	2
Parent/offspring	1264	1264	3792
Sib/Sib	948	948	948
Grandparent	438	0	0
Avuncular	469	0	0
Half sib	28	0	0
Cousin	106	0	0
Individuals (Male/Female)	940(502/436/2)	1042(565/477)	3792(2007/1785)
Founder/Non founder/Singleton	306/632/2	410/632/0	1896/1896

NOTE. – Different data structures:

* ORIGINAL: initial data set in multigenerational family structure

|| ONE: one parental genotype was removed at each marker locus in top and middle generations

 Δ TOP: both founder genotypes were removed at each marker locus in top generation

| TWO: all parental genotypes were removed at each marker locus in top and middle generations

† NUCLEAR: nuclear families without two parental genotypes at all marker loci

‡ SIBPAIR: single sib pairs without parental genotypes at all marker loci in nuclear family framework.

TABLE II
Observed LOD scores and p-values for 20 traits – 2 cM and 0.25 cM density of SNP markers, as proportion of missing parental genotypes and pedigree structure change

Data sets		ORIGINAL*			ONE ^{††}			TOP ^Δ			TWO [‡]			NUCLEAR [†]			SIBPAIR ^{‡‡}		
Trait	Density of SNP markers	2cM	0.25cM	Chr	2cM	LOD	P-***	2cM	LOD	P-***	2cM	LOD	P-***	2cM	LOD	P-***	2cM	LOD	P-***
Drs1796969	0.36	3.73	3.66	1	4.49	3.71	<0.0001	4.23	2.77	0.0002	4.7	2.56	0.0011	2.04	2.08	0.0010	1.28	2.01	0.0012
Disc0583756	0.35	1.68	2.19	2	1.87	3.10	0.0020	1.88	2.91	<0.0001	2.21	2.24	0.0007	2.32	2.12	0.0009	1.1	1.28	0.0080
Drs269384	0.12	6.03	7.65	3	5.84	8.06	<0.0001	5.83	7.69	<0.0001	5.54	7.41	0.0001	5.94	7.05	<0.0001	5.98	13.27	<0.0001
Drs1051447	0.30	6.11	6.62	4	6.79	6.92	<0.0001	5.46	5.12	<0.0001	4.15	4.56	<0.0001	3.96	3.19	<0.0001	3.7	4.85	<0.0001
Drs34999	0.36	3.33	4.17	5	3.52	4.29	<0.0001	3.24	4.04	<0.0001	2.3	3.22	0.0001	1.12	2.14	0.0009	0.39	1.85	0.0020
Drs159988	0.18	4.73	7.22	6	5.26	6.98	<0.0001	5.44	6.31	<0.0001	5.18	5.19	<0.0001	3.05	3.11	0.0001	1.46	4.23	<0.0001
Drs42611	0.38	3.4	4.26	7	3.3	3.90	0.0001	3.29	3.63	<0.0001	3.06	2.63	0.0002	1.48	2.38	0.0005	0.75	1.21	0.0090
Drs765262	0.30	6.81	6.75	8	5.27	5.89	<0.0001	4.06	5.07	<0.0001	3.04	4.35	0.0001	1.95	1.89	0.0020	0.13	1.77	0.0020
Drs363717	0.20	9.54	13.03	9	9.18	13.23	<0.0001	7.07	11.37	<0.0001	7.05	10.16	<0.0001	6.26	8.66	<0.0001	3.42	5.94	<0.0001
Disc0040918	0.31	8.51	8.66	10	7.69	8.51	<0.0001	6.74	7.57	<0.0001	6.1	7.26	<0.0001	4.96	5.12	<0.0001	2.45	5.11	<0.0001
Disc0912310	0.18	8.01	9.36	11	8.39	8.88	<0.0001	7.38	7.52	<0.0001	6.88	6.58	<0.0001	4.64	4.24	<0.0001	3.2	6.56	<0.0001
Disc0591807	0.22	6.21	7.08	12	6.51	7.14	<0.0001	5.78	6.69	<0.0001	7.18	6.09	<0.0001	3.99	6.26	<0.0001	2.31	4.87	<0.0001
Disc0054081	0.25	4.54	5.21	13	4.15	5.30	<0.0001	3.71	4.47	<0.0001	3.65	3.71	<0.0001	2.76	3.70	<0.0001	1.48	4.05	<0.0001
Disc0914256	0.25	3.93	4.49	14	3.73	4.60	<0.0001	3.97	3.56	<0.0001	3.84	3.14	0.0001	2.78	2.61	0.0003	3.93	7.76	<0.0001
Drs488756	0.15	4.37	4.18	15	4.78	3.47	<0.0001	3.44	2.80	0.0002	3.89	2.68	<0.0001	3.87	3.26	0.0001	7.5	7.40	<0.0001
Disc0513689	0.31	5.38	7.99	16	5.89	7.41	<0.0001	5.38	7.30	<0.0001	5.55	6.47	<0.0001	3.5	3.28	<0.0001	3.57	5.02	<0.0001
Drs757288	0.20	8.05	8.89	17	8.21	9.36	<0.0001	6.22	7.55	<0.0001	5.4	6.09	<0.0001	6.35	7.90	<0.0001	5.67	10.03	<0.0001
Drs1941207	0.30	4.35	5.71	18	3.72	5.28	<0.0001	2.71	3.72	<0.0001	1.89	3.62	<0.0001	2.48	2.74	<0.0001	2.53	3.33	<0.0001
Drs1603	0.24	12.29	12.03	19	11.82	11.92	<0.0001	10.25	11.26	<0.0001	8.67	10.25	<0.0001	6.69	7.11	<0.0001	12.77	13.08	<0.0001
Disc0046507	0.29	4	4.96	20	4.48	4.50	<0.0001	4.43	4.08	<0.0001	4.15	4.26	<0.0001	5.46	5.06	<0.0001	3.06	3.51	<0.0001

NOTE. – **Underlined bold:** The highest LOD score in each trait at each density level, *Italic:* LOD score less than 2.

Chr: Chromosomal location of the marker used to create the trait

LOD: maximum Kong & Cox LOD score within 20 cM of the true trait locus

Different data structures:

- #_{MAF}: Minor Allele Frequency of risk allele used to create the trait
- ** p value: minimum p value of NPL score statistics within 20 cM of the true trait locus
- * ORIGINAL: initial data set in multigenerational family structure
- $\int \int$ ONE: one parental genotype was removed at each marker locus in top and middle generations
- Δ TOP: both founder genotypes were removed at each marker locus in top generation
- $\int$ TWO: all parental genotypes were removed at each marker locus in top and middle generations
- $\dagger$ NUCLEAR: nuclear families without two parental genotypes at all marker loci
- # SIBPAIR: single sib pairs without parental genotypes at all marker loci in nuclear family framework.

TABLE III

False positive frequencies for 1,000 simulated traits

Data set	Marker set	Number of SNPs	Number of false positive below p-value criterion of					
			p<0.05	p<0.01	p<0.0017	p<0.001	p<0.0001	p<0.000049
ORIGINAL	0.25 cM	15019	24271	6806	1512	997	123	60
	0.3 cM	5405	17474	4887	924	613	74	31
	0.6 cM	3671	14213	3857	880	590	90	49
	1 cM	2625	13888	3949	827	557	77	42
	2 cM	1539	12806	3899	816	545	74	37
ONE	0.25 cM	15019	24047	7022	1625	1132	135	77
	0.3 cM	5405	16226	4654	913	581	73	37
	0.6 cM	3671	14852	4367	887	578	69	29
	1 cM	2625	12986	3728	724	481	73	43
	2 cM	1539	11446	3101	533	346	44	23
TOP	0.25 cM	15019	24835	8042	2263	1673	343	205
	0.3 cM	5405	15098	4122	770	473	46	23
	0.6 cM	3671	13921	3767	693	450	42	21
	1 cM	2625	12487	3334	602	371	47	23
	2 cM	1539	10020	2495	459	295	27	16
TWO	0.25 cM	15019	24545	8276	2588	1924	452	272
	0.3 cM	5405	14587	4279	837	556	49	21
	0.6 cM	3671	13143	3814	861	581	93	42
	1 cM	2625	11652	3321	717	481	72	39
	2 cM	1539	11196	3253	651	425	43	18
NUCLEAR	0.25 cM	15019	25527	8446	2472	1831	388	276
	0.3 cM	5405	14152	4254	1054	741	109	63
	0.6 cM	3671	12931	3911	909	623	81	43
	1 cM	2625	11110	3109	627	440	45	25
	2 cM	1539	9614	2623	478	295	43	17
SIBPAIR	0.25 cM	15019	41498	18725	8180	6725	2631	2015
	0.3 cM	5405	15143	5394	1562	1139	220	111
	0.6 cM	3671	14498	5143	1507	1087	204	113
	1 cM	2625	13285	4759	1465	1069	228	136
	2 cM	1539	13597	5212	1744	1338	318	211

NOTE. - Different data structures:

* ORIGINAL: initial data set in multigenerational family structure

$\int \int$ ONE: one parental genotype was removed at each marker locus in top and middle generations

Δ TOP: both founder genotypes were removed at each marker locus in top generation

$\int \int$ TWO: all parental genotypes were removed at each marker locus in top and middle generations

$\dagger$ NUCLEAR: nuclear families without two parental genotypes at all marker loci

$\#$ SIBPAIR: single sib pairs without parental genotypes at all marker loci in nuclear family framework.