Multiple Subsampling of Dense SNP Data Localizes Disease Genes with Increased Precision

William C.L. Stewart Anna L. Peljto David A. Greenberg

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, N.Y., USA

Key Words

Linkage disequilibrium sequencing · Subsampling · Candidate gene · Linkage analysis

Abstract

Background/Aims: Current linkage studies detect and localize trait loci using genotypes sampled at hundreds of thousands of single nucleotide polymorphisms (SNPs). Such data should provide precise estimates of trait location once linkage has been established. However, correlations between nearby SNPs can distort the information about trait location. Traditionally, when faced with this dilemma, three approaches have been used: (1) ignore the correlation; (2) approximate the correlation; or, (3) analyze a single, approximately uncorrelated subset of the original dense data. **Methods:** Here, we examine and test a simple and efficient estimator of trait location that averages location estimates across random subsamples of the original dense data. Based on pairwise estimates of correlation, we ensure that the SNPs within each subsample are approximately uncorrelated. In addition, we use the nonparametric bootstrap procedure to compute narrow, high-resolution candidate gene regions (i.e. confidence intervals for the true trait location). **Results:** Using simulated data, we show that the three existing approaches to dense SNP linkage analysis (described above) can yield biased and/or inefficient estimation depending on the un-

derlying correlation structure. With respect to mean squared error, our estimator outperforms the third approach, and is as good as, but usually better than the first and second approaches. Relative to the third approach, our estimator led to a 47.5% reduction in the candidate gene region length based on the analysis of 15 hypertension families genotyped at ~500,000 SNPs. **Conclusion:** The method we developed will be an important tool for constructing high-resolution candidate gene regions that could ultimately aid in targeting regions for sequencing projects.

Copyright © 2009 S. Karger AG, Basel

Introduction

In genetic studies of complex traits, like type 2 diabetes, Chron's disease, and idiopathic generalized epilepsy, the initial identification of chromosomes that harbor disease-related genes is often first established by genome-wide linkage analysis [Greenberg et al., 2000; Hugot et al., 2001; Grant et al., 2006; Pal et al., 2008]. Despite the fact that most modern linkage studies collect genotypes at hundreds of thousands of single nucleotide polymorphisms (SNPs), these studies are rarely successful at fine-mapping disease genes by association alone. In fact, once linkage to a broad region has been established, using the dense SNP data to extract additional information about

inheritance is often the most effective way to improve disease gene localization. To this end, we describe an efficient, multiple subsampling approach that uses the dense SNP data of modern linkage studies to produce accurate, narrow confidence intervals for the locations of disease-related genes. Furthermore, our confidence intervals may be interpreted as candidate gene regions (CGRs) that could then inform subsequent candidate gene and/or sequencing studies.

Since dense SNP linkage data have been shown to carry considerable information about linkage [Evans and Cardon, 2004], CGRs constructed from such data are expected to be accurate and precise. However, dense SNP data tend to exhibit correlation between SNPs, and this can negatively affect the precision and accuracy of CGRs. The correlation between SNPs is known as linkage disequilibrium (LD), and existing multipoint linkage methods either (1) ignore LD; (2) crudely approximate LD; or (3) analyze a single, approximately uncorrelated subset of the original dense data. In any event, LD is either ignored or poorly modeled, or the bulk of the dense SNP genotype information is ignored.

All three strategies have their own advantages, disadvantages, and unknowns. For example, while it is well-known that ignoring LD leads to false positives in the absence of linkage [Huang et al., 2004; Abecasis and Wiginton, 2005; Cho et al., 2007; Li and Leal, 2008], much less is known about the effect of ignoring LD when linkage exists. Intuitively, since the pedigree likelihood is incorrectly specified, one might expect biased results with increased variability.

As another example, consider methods that account for LD. Typically these methods transform the original dense SNP data with the aim of improving the fit of the standard pedigree likelihood. In particular, the genetic analysis package MERLIN [Abecasis et al., 2002] transforms the original dense SNP data by clustering adjacent SNPs that are highly correlated. Within each cluster, the string of contiguous genotypes are assumed to originate from a small number of non-recombining haplotypes. Thus, the non-recombining haplotypes of a cluster are the alleles of that cluster, and each cluster is simply a highly polymorphic marker. MERLIN then assumes independence between the clusters, which is equivalent to assuming the absence of LD. The utility of this approach depends crucially on the assumption that the human genome is subdivided into short, independent 'LD-blocks' [Daly et al., 2001; Goldstein, 2001; Reich et al., 2001]. However, this assumption is highly questionable for recently admixed populations (e.g. African Americans) [Lewont-

in, 1974; Briscoe et al., 1994], genetic isolates (e.g. Old Order Amish) [Puffenberger, 2003], and for specific regions of the genome.

As another example of a program that takes LD into account, the linkage program SNPLINK [Webb et al., 2005] avoids any potentially negative effects of LD by using a set of approximately uncorrelated SNPs, which are usually sparsely spaced. However, this strategy ignores the vast majority of the original dense SNP data, and as such, some information is necessarily lost. Since the likelihood under the correct model is intractable, quantifying the amount of information lost when using a particular subset of the original dense SNPs is difficult. As such, it generally is not possible to determine an optimal set of approximately uncorrelated SNPs.

In this study, we propose a subsampling-based approach that makes efficient use of the original dense SNP data to accurately estimate the true trait location. In our approach, we repeatedly subsample the original dense SNP data, but for each subsample we retain the original families and their phenotypic information. The subsamples are chosen at random, subject to the constraint that the SNPs are approximately uncorrelated. Then, for each subsample we compute the maximum likelihood (ML) estimate of trait location, and average these ML estimates across subsamples to obtain a more accurate and efficient estimate of trait location. Our estimator is consistent and asymptotically normally distributed; and, when used in conjunction with the nonparametric bootstrap procedure [Efron and Tibshirani, 1993], it yields asymptotically correct confidence intervals which may be interpreted as CGRs. By theoretical arguments and by simulation, we show that our estimator has lower mean squared error (MSE) than existing methods. We not only test our method using extensive simulation, but we also show its utility by applying it to the dense SNP linkage data of 15 pulmonary hypertension families [Rodriguez-Murillo et al., accepted]. Our approach is implemented in the program: EAGLET (Efficient Analysis of Genetic Linkage: Estimation and Testing), and is freely available to the public (see Web References).

Methods

Estimation with Subsampling

For an arbitrary pedigree, let the pair (T, G) denote all trait and genotype data observed on some (possibly all) members of the pedigree across N SNPs. Define $\mathbf{G} \equiv (G_1, \dots, G_N)$ as the collection of single marker genotypes, where G_i denotes the marker-specific genotypes observed at the i -th SNP for $i = 1, 2, \dots, N$. For ease of

exposition, we describe our estimator in the context of completely parametric linkage analysis; however, theoretical and empirical results for trait model-free linkage methods are discussed later.

Traditionally, estimation of trait location requires maximization of the pedigree likelihood

$$\begin{aligned} L(\xi; \mathbf{t}, \mathbf{g}, \lambda) &\equiv P_{\xi}(\mathbf{T}, \mathbf{G}; \lambda), \\ &\propto P_{\xi}(\mathbf{G} | \mathbf{T}, \lambda) \end{aligned} \quad (1)$$

where λ specifies all parameters related to penetrance, haplotype frequencies, and to the genetic map, and ξ is a univariate parameter denoting the location of the trait locus relative to the known locations of the N SNPs. In the absence of LD, a variety of different algorithms [Elston and Stewart, 1971; Baum, 1972; Lander and Green, 1987; Irwin et al., 1994; Thompson, 2000] permit the computation or consistent estimation of the pedigree likelihood for a given value of ξ . However, in the presence of LD, exact computation and/or consistent estimation of the pedigree likelihood is usually not feasible. Therefore, to estimate trait location in the presence of LD, the standard approach uses the maximizer of

$$L(\xi; \mathbf{g}, \mathbf{t}, \lambda) \propto P_{\xi}(\mathcal{G} | \mathbf{T}; \lambda), \quad (2)$$

where $\mathcal{G} \equiv \{G_{ij}\}$ is a subsample of $\mathbf{G}$ that retains the genotypes at $M < N$ approximately uncorrelated SNPs for all individuals in the original dense data. An approximately uncorrelated subsample has $0 \leq r^2 < t$ for all pairs of adjacent SNPs in the subsample for some preset value of t . In particular,

$$\tilde{\xi}(\mathcal{G}) \equiv \operatorname{argmax} \{P_{\xi}(\mathcal{G}; \lambda)\} \quad (3)$$

is the ML estimate of ξ given $\mathcal{G}$. Since there is a large number of approximately uncorrelated subsamples from which to choose, there is no obvious way of knowing which subsample yields the most accurate estimate of ξ .

We propose the following new and improved estimator of trait location that makes use of all available data in $\mathbf{G}$ by combining information across multiple subsamples. Specifically, for K random subsamples $(\mathcal{G}_1, \dots, \mathcal{G}_K)$, with $\mathcal{G}_u$ independent of $\mathcal{G}_v$, for $u \neq v$, define

$$\begin{aligned} \hat{\xi}(\mathbf{G}) &\equiv \mathbf{E}(\tilde{\xi}) \\ &\approx \frac{1}{K} \sum_{u=1}^K \tilde{\xi}(\mathcal{G}_u). \end{aligned} \quad (4)$$

To obtain K random subsamples, we first use the expectation-maximization (EM) algorithm [Dempster et al., 1977] to obtain pairwise estimates of LD for all adjacent SNPs. Then, based on a minimum level of heterozygosity and on a maximal level of pairwise LD, we sample informative and approximately uncorrelated subsets of the original dense SNP data. Note that for notational convenience, the dependence of all trait location estimators on trait data $\mathbf{T}$ is suppressed.

Theoretical Properties

To facilitate the exposition of theoretical properties of our estimator, a slight reformulation of the problem is useful. Suppose that there are $R > 1$ approximately uncorrelated, distinct subsamples. For a given subsample, let the binary vector of SNP indicators $\mathbf{S}$ denote the corresponding subset of SNPs, and let $P_{\lambda}(\mathbf{S}; t)$ be a non-degenerate probability distribution over the space of bi-

nary vectors $(\mathbf{S}_1, \dots, \mathbf{S}_R)$. With this notation, a subsample is alternatively defined by the function $h(\mathbf{G}, \mathbf{S})$, where h retains only the marker-specific genotypes in $\mathbf{G}$ indicated by the 1's in $\mathbf{S}$. From this vantage point, it is easier to see that the standard estimate of trait location, $\tilde{\xi}(\mathcal{G})$ defined in (3), is random in both $\mathbf{G}$ and $\mathbf{S}$; whereas our estimator $\hat{\xi}$ defined in (4) is only random in $\mathbf{G}$. As a result, our estimator $\hat{\xi}$ has smaller variance than $\tilde{\xi}$ since

$$\begin{aligned} \operatorname{Var}(\hat{\xi}) &\equiv \operatorname{Var}_{\mathbf{G}}(\mathbf{E}_{\mathbf{S}} \tilde{\xi}) \\ &\leq \operatorname{Var}_{\mathbf{G}}(\mathbf{E}_{\mathbf{S}} \tilde{\xi}) + \mathbf{E}_{\mathbf{G}} \operatorname{Var}_{\mathbf{S}}(\tilde{\xi}) \\ &= \operatorname{Var}(\tilde{\xi}). \end{aligned} \quad (5)$$

Furthermore, the magnitude of the reduction is equal to the average Monte Carlo variance in $\tilde{\xi}$. The subscripts on the functionals in (5) indicate the marginal distributions over which expectations and variances are taken.

To establish the asymptotic consistency of $\hat{\xi}$, we need to consider genotype and trait data $\mathbf{G}_n$ and $\mathbf{T}_n$, respectively, observed on members of n independent and identically distributed (i.i.d.) families (e.g. affected sibling pairs (ASPs)). As before, the dependence on trait data is suppressed. For dense SNP data $\mathbf{G}_n$ and for fixed $j \in \{1, \dots, R\}$, let $\xi_{nj} \equiv \tilde{\xi}(h(\mathbf{G}_n, \mathbf{S}_j))$ denote the ML estimator of ξ . Provided that the usual regularity conditions hold, ξ_{nj} is consistent for the true trait location ξ^* as $n \rightarrow \infty$. Now, given the consistency of ξ_{nj} for each j , the functional form of (4) implies the consistency of $\hat{\xi}$ as $n \rightarrow \infty$. To establish the asymptotic normality of $\hat{\xi}$, let $Y_{nj} \equiv \sqrt{n}(\xi_{nj} - \xi^*)$, so that Y_{nj} is asymptotically normal with mean zero and variance σ_j^2 which in general will depend on $\mathbf{S}_j$. Now, $\sqrt{n}(\hat{\xi}_n - \xi^*) = \sum_{j=1}^R w_j \cdot Y_{nj}$, where $w_j = P_{\lambda}(\mathbf{S}_j; t)$. Hence, by Slutsky's theorem, each term in the sum is asymptotically normal with mean zero and possibly different variances, and by the Lindeberg-Feller central limit theorem, $\hat{\xi}_n$ is asymptotically normal.

Data Description

To assess the performance of the different methods that account for LD in the presence of linkage, we analyzed 500 replicates of simulated, dense SNP linkage data for 28 different settings. We also analyzed the real data of 15 pulmonary hypertension families to compare our method to the commonly used approach that constructs CGRs from a single, random subsample. To simulate dense SNP linkage data in the presence of LD, we used the program Caleb (see Web References). The details of the different settings are discussed below.

For all settings, we simulated a dichotomous trait in the middle of a 49-cM map with 98 equi-spaced SNPs in LD (average spacing 0.5 cM). The trait had 50% penetrance with a prevalence of 9%, and was not in LD with any of the SNPs. The phenotypic and genotypic data of founders were masked, and allele frequencies were fixed on the basis of estimates obtained from the real data of a dense SNP linkage scan. Each setting differed in at least

Fig. 1. In all four plots, horizontal lines indicate the level of LD between pairs of adjacent SNPs, and all four LD patterns were used to simulate linkage data. The LD pattern shown in the lower righthand corner was estimated from real data, while the other patterns were chosen as instructive examples to characterize the effects of different patterns of extreme LD.

one of the following: LD pattern (we explored four different LD patterns), disease model (dominant or recessive), and power to detect linkage (expected lod (ELOD) equal to 2.0, 2.5, or 3.0).

Our simulations used four different LD patterns. For the first and second LD patterns we simulated contiguous blocks of LD. Within each block and within each interval, we fixed the value of D' and computed the average r^2 over consecutive intervals for comparison. The first LD pattern had two equi-length blocks: one with weak LD ($D' = 0.1$; average $r^2 = 0.0037$), the other with strong LD ($D' = 0.9$; average $r^2 = 0.3106$). The second had three equi-length blocks of alternating LD (i.e. strong ($D' = 0.9$; average $r^2 = 0.2997$), weak ($D' = 0.1$; average $r^2 = 0.0034$), and strong ($D' = 0.9$; average $r^2 = 0.3392$)). The third LD pattern is similar to the second, except that the LD is allowed

to vary continuously along the chromosome (average $r^2 = 0.1885$; see fig. 1). The fourth LD pattern was estimated from the real data of a dense SNP linkage scan (average $r^2 = 0.1448$).

In addition to the ASP families, we also examined the effect of using larger nuclear families with two affected and one unaffected (AAU) offspring, with three affected (AAA) offspring, and with a mixture of ASP, AAU, and AAA families. Lastly, we applied our method in conjunction with the nonparametric bootstrap procedure to the dense SNP data of 15 pulmonary hypertension families. We restricted attention to the CGR centered at chromosome 3p12 [Rodriguez-Murillo et al., accepted]. For all lod score calculations and subsequent estimation of trait location, we used the Z_{lr} linkage statistic [Kong and Cox, 1997].

Table 1. CGR lengths for ASPs based on EAGLET

	LD-1	LD-2	LD-3	LD-4
DOM				
ELOD = 2.0	30 (93)	26 (93)	24 (94)	28 (93)
ELOD = 2.5	25 (94)	25 (93)	23 (93)	24 (95)
ELOD = 3.0	21 (93)	22 (94)	21 (93)	23 (94)
REC				
ELOD = 2.0	28 (92)	26 (93)	25 (93)	27 (93)
ELOD = 2.5	22 (93)	22 (93)	22 (94)	27 (93)
ELOD = 3.0	22 (94)	20 (93)	20 (94)	25 (93)

Candidate gene region lengths given in cM (i.e. the lengths of the 95% confidence intervals) based on EAGLET are shown. Estimated coverage probabilities based on 500 replicates are given in parentheses. For each LD pattern 1–4 (described in the text), and for each genetic model, dominant (DOM) and recessive (REC), three levels of the ELOD are considered.

Table 2. MSE for ASPs with ELOD = 2.0

	EAGLET	MERLIN	SNPLINK	Equi-Distant
LD-1				
DOM	57	70	108	74
REC	50	74	102	63
LD-2				
DOM	45	66	69	76
REC	44	79	81	84
LD-3				
DOM	39	78	85	77
REC	42	77	91	78
LD-4				
DOM	50	71	84	80
REC	49	79	91	62

The estimated MSE for the different methods that account for LD is shown across LD patterns 1–4 (described in the text), and across genetic models: dominant (DOM) and recessive (REC). In each row, the lowest MSE is given in bold.

Results

In the context of ASPs, we evaluated four different methods that account for LD. We estimated the MSE of trait location estimators, and their corresponding CGRs and coverage probabilities. We used the following four approaches: (1) a single subsample with equi-distant SNPs (denoted Equi-Distant in what follows), (2) the multi-

point linkage program MERLIN, (3) SNPLINK, and (4) our estimator: EAGLET. Recall that EAGLET averages trait location estimates over multiple subsamples of the original dense SNP linkage data to estimate the true trait location, and uses the nonparametric bootstrap procedure to construct CGRs. From table 1, we see that EAGLET has good coverage across a wide range of ASP designs, and that CGR lengths appear to decrease as the ELOD increases for each LD pattern. In addition, the table shows that within each row, the estimated CGR lengths are approximately constant, suggesting that EAGLET estimates are not influenced by LD. The results (table 2), show that EAGLET also has the lowest MSE of all four methods. Table 2 also shows that for each row, the MSEs of MERLIN, SNPLINK, and Equi-Distant are closer to one another than they are to the MSE of EAGLET. This apparent clustering may arise from the fact that MERLIN, SNPLINK, and Equi-Distant only use a single, random cluster or subsample when estimating trait location, whereas EAGLET makes use of all of the available dense SNP data by using multiple random subsamples.

Since the bias is relatively small across all settings for all four methods (data not shown), the uniform increase in efficiency for EAGLET (see table 2) implies a corresponding decrease in the length of the candidate gene region (table 3). This general pattern holds for the other levels of the ELOD as well (data not shown). By contrast, when LD is ignored in the analysis and the asymmetric LD pattern (LD-1) is used to simulate the data, ASPs with a dominant trait yield a bias of 2.07 cM in the direction of the strong ($D' = 0.9$) LD block. In fact, a similar result holds for every setting where the asymmetric pattern of LD was used for simulation but LD was ignored in the analysis, irrespective of the disease model (dominant or recessive) or of the power (ELOD = 2.0, 2.5, and 3.0).

However, consistent with the findings of Abecasis and Wigginton [2005] and Huang et al. [2004], ignoring LD has negligible effect on trait location estimation when parental genotypes are available (data not shown).

Also, in the context of larger nuclear families, we compared the four methods that account for LD (EAGLET, MERLIN, SNPLINK, and Equi-Distant) with the method that ignores LD. The family structures we considered were AAA and AAU. Table 4 shows that, compared to the other methods, the MSE of EAGLET does not change much between AAA and AAU settings for either genetic model. Here, the AAA and AAU settings each use the continuously varying pattern of LD, with an ELOD of 2.0. Similar results were obtained for the other AAA and AAU settings that use different LD patterns and different

sample sizes (data not shown). Recall that Huang et al. [2004] showed that genotyping of additional unaffected family members reduces the negative impact of LD. Therefore, for methods that account well for LD, there will be very little change in MSE between AAA and AAU designs. But, for methods that account poorly (or not at all) for LD, there will be considerable change in MSE between AAA and AAU designs. Hence, the results in table 4 provide additional evidence that, relative to MERLIN and SNPLINK, EAGLET better accounts for LD.

When analyzing real data, an estimate of the variance of our estimator is needed to construct 95% confidence intervals or CGRs. When data come from unrelated families, an effective way to estimate the variance of our estimator is through the nonparametric bootstrap procedure [Stewart, 2007]. To evaluate the utility of the bootstrap, we simulated 500 replicates of dense SNP linkage data using the empirical LD pattern (LD-4) with a dominant model. For each replicate we used the nonparametric bootstrap in conjunction with our estimator-of-trait location to construct a CGR. Across the 500 CGRs constructed in this way, we computed the coverage probability for settings where the ELOD was 2.0, 2.5, and 3.0. The corresponding coverage probabilities were: 94.8, 96.8, and 94.8 percent, respectively. Each replicate contained a mixture of ASP, AAA, and AAU families. Finally, to evaluate the utility of our proposed method in a real-data application, we re-estimated the CGR on chromosome 3p12 using the dense SNP linkage data of 15 pulmonary hypertension families. The re-estimated CGR length based on EAGLET is 9.02 cMs, which is 47.5% shorter than the estimated CGR based on the standard (i.e. Equi-Distant) approach.

Discussion

In this work, we have shown that our proposed estimator of trait location efficiently uses the original dense SNP data to minimize the negative effects of LD over a range of LD patterns, pedigree structures, and genetic models. We have demonstrated that ignoring LD when it is present leads to bias in estimates of trait location and increased variance; and that relative to existing methods that account for LD, our estimator has the smallest MSE with the sharpest CGRs while maintaining approximate 95% nominal coverage. What may be the most useful and surprising aspect of our proposed method is its extreme generality. In principle, our subsampling approach is easily integrated with any existing linkage program, which could increase the resolution of CGRs for investigators

Table 3. CGR lengths for ASPs with ELOD = 2.0

	EAGLET	MERLIN	SNPLINK	Equi-Distant
LD-1				
DOM	30 (93)	33 (93)	41 (91)	34 (90)
REC	28 (92)	34 (92)	39 (91)	31 (92)
LD-2				
DOM	26 (93)	32 (92)	31 (92)	34 (91)
REC	26 (93)	35 (92)	32 (93)	36 (92)
LD-3				
DOM	24 (94)	35 (92)	31 (90)	34 (92)
REC	25 (93)	34 (91)	35 (89)	35 (89)
LD-4				
DOM	28 (93)	33 (92)	36 (92)	35 (91)
REC	27 (93)	35 (92)	37 (91)	31 (92)

For each of the methods that account for LD, the corresponding candidate gene region lengths given in cM (i.e. the lengths of the 95% confidence intervals) are shown. LD patterns 1–4 (described in the text) and the genetic models, dominant (DOM) and recessive (REC), are considered.

Table 4. MSE for AAA and AAU families with LD-3 and ELOD = 2.0

	EAGLET	MERLIN	SNPLINK	Equi-Distant	Ignoring LD
DOM					
AAA	43	89	79	77	97
AAU	41	62	64	60	77
REC					
AAA	50	95	101	75	126
AAU	38	64	62	65	73

Estimated MSE is shown for five different linkage analysis methods. The first four methods account for LD, while the fifth method completely ignores LD. The data were simulated using the LD-3 pattern described in the text. Results are shown for dominant (DOM) and recessive (REC) genetic models, and for two types of nuclear families with unobserved parental data: three affected offspring (AAA), and two affected offspring with an unaffected sibling (AAU). In each row, the lowest MSE is given in bold.

with dense SNP linkage data on large, extended pedigrees in the presence of LD. An excellent resource for linkage programs that are designed for large, extended pedigrees is available through the MORGAN package (see Web References).

Another computational advantage of our approach stems from the interplay between theory and implementation. Efron and Tibshirani (1993) noted that the number of nonparametric bootstrap resamples needed to accurately estimate the variance of a statistic of interest is markedly less than the number needed to accurately estimate the quantiles of the corresponding distribution. By establishing the asymptotic normality of our statistic, we only need estimate its variance in order to construct valid confidence intervals. Thus, relative to the computation time required for the standard bootstrap confidence interval, which is based on the quantiles, the computation time needed to construct our confidence intervals is dramatically decreased.

Fundamentally, under the assumption of random mating, the probability that affected relatives share alleles identical by descent is influenced either by chance alone, or both by chance and selection. If an allele at a locus confers increased risk for disease, then affected relatives will have a higher probability of sharing this allele than would be expected by chance alone. However, this can also occur at markers for which the alleles do not confer differential risk, provided that the LD in the population is ignored in the analysis, and that the founders are not observed. In this case, the apparent sharing among affected relatives will exceed the level predicted by chance alone, and this will provide false evidence for linkage. If this false evidence for linkage is strong relative to the true evidence for linkage, then inefficient estimation will result when LD is ignored and/or poorly modeled. As demonstrated by the analysis of real data, typical patterns of LD are sufficient to bring about non-ignorable levels of increased variability in estimates of trait location. Furthermore, the results of the simulated data show that (1) relatively long stretches of LD (which can occur in recently admixed populations, or genetic isolates) may be particularly problematic; and, (2) our method is expected to yield considerable gains in accuracy and/or efficiency in such settings.

EAGLET and MERLIN manipulate the original dense SNP data in fundamentally different ways: EAGLET draws approximately uncorrelated subsamples; whereas, MERLIN imputes LD-blocks. In addition, there is considerable scope within each method to adjust the extent to which the original dense data are manipulated. For example, when different thresholds for the maximum pairwise r^2 are used in conjunction with EAGLET, the resultant subsamples will typically contain different numbers of SNPs. Therefore, to ensure that the differences in performance (if any) are due to differences in methodology

we need to control the extent to which we allow each method to manipulate the original dense SNP data. Since all of the methods use the standard pedigree likelihood from equation (2), and since this likelihood assumes the absence of LD between markers, it makes sense to standardize the four methods that account for LD by fixing the average number of markers considered. Then, any observed differences must be due to the different methodologies, and cannot be due to the user-defined inputs that govern the extent to which the original dense SNP data are manipulated. We decided to fix the average number of markers at 24 since this agrees with the recommendation of Abecasis and Wigginton [2005] in the sense that an average of 24 SNPs per EAGLET subsample corresponds to an r^2 threshold of about 0.10.

Similarly, the average number of SNPs for SNPLINK and Equi-Distant were also fixed at 24, as was the average number of MERLIN clusters.

Historically, the negative effect of LD on efforts to map disease genes through linkage analysis has been largely ignored. However, an indirect consequence of the widespread use of high-throughput genotyping technology is that opportunities for improved linkage methods that account for LD abound. In addition, the limited success of genome-wide association studies seems to underscore the need for improved linkage methods in general. Our proposed method addresses this need, and permits the rapid construction of accurate, high-resolution candidate gene regions from dense SNP linkage data. Thus, investigators now have an extremely useful and reliable tool for localizing disease susceptibility genes.

We are in the process of extending our multiple subsampling approach to the task of linkage detection, where we expect to achieve similar gains in power.

Acknowledgements

We would like to thank Esther Drill, Adam Greenberg, and Junying Zhang for their invaluable computer support. We would also like to acknowledge the support of NIH grants NHLBI NLO60056, Training Grant, NIDDK DKO31175, DKO31813, and the Professional Schools Fund of Columbia University.

References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 2002;30:97–101.
- Abecasis GR, Wigginton JE: Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *American Journal of Human Genetics* 2005;77:754–767.
- Baum L: An inequality and associated maximization technique in statistical estimation of probabilistic functions of markov processes. *Inequalities* 1972;3:1–8.
- Briscoe D, Stephens JC, O'Brien SJ: Linkage disequilibrium in admixed populations: applications to gene mapping. *Journal of Heredity* 1994;85:59–63.
- Cho K, Yang Q, Dupuis J: Handling linkage disequilibrium in linkage analysis using dense single-nucleotide polymorphisms. *BMC Proceedings* 2007;(suppl 1):S161.
- Daly MJ, et al: High-resolution haplotype structure in the human genome. *Nature Genetics* 2001;29:229–232.
- Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977;39:1–37.
- Efron B, Tibshirani RJ: *An Introduction to the Bootstrap*. New York, Chapman & Hall, 1993.
- Elston RC, Stewart J: A general model for the analysis of pedigree data. *Human Heredity* 1971;21:523–542.
- Evans DM, Cardon LR: Guidelines for genotyping in genomewide linkage studies: single-nucleotide polymorphism maps versus microsatellite maps. *American Journal of Human Genetics* 2004;75:687–692.
- Goldstein DB: Islands of linkage disequilibrium. *Nature Genetics* 2001;29:109–111.
- Grant SF, et al: Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics* 2006;38:320–323.
- Greenberg DA, et al: Reproducibility and complications in gene searches: linkage on chromosome 6, heterogeneity, association, and maternal inheritance in juvenile myoclonic epilepsy. *American Journal of Human Genetics* 2000;66:508–516.
- Huang Q, Shete S, Amos CI: Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *American Journal of Human Genetics* 2004;75:1106–1112.
- Hugot JP, et al: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature Genetics* 2001;41:599–603.
- Irwin M, Cox N, Kong A: Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* 1994;91:11684–11688.
- Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics* 1997;61:1179–1188.
- Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* 1987;84:2363–2367.
- Lewontin RC: *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York, N.Y., 1974.
- Li B, Leal SM: Ignoring intermarker linkage disequilibrium induces falsepositive evidence of linkage for consanguineous pedigrees when genotype data is missing for any pedigree member. *Human Heredity* 2008;65:199–208.
- Pal DK, Strug LJ, Greenberg DA: Evaluating candidate genes in common epilepsies and the nature of evidence. *Epilepsia* 2008;49:2258–2265.
- Puffenberger EG: Genetic heritage of the Old Order Mennonites of southeastern Pennsylvania. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 2003;121C:18–31.
- Reich DE, et al: Linkage disequilibrium in the human genome. *Nature* 2001;41:199–204.
- Rodriguez-Murillo, et al: (Accepted). Novel loci interacting epistatically with Bone Morphogenetic Protein Receptor 2 cause familial pulmonary arterial hypertension. *Journal of Heart and Lung Transplantation*.
- Stewart WCL: Improving estimates of genetic maps: A meta-analysis based approach. *Genetic Epidemiology* 2007;31:408–416.
- Thompson EA: *Statistical Inferences from Genetic Data on Pedigrees*. NSFCBMS Regional Conference Series in Probability and Statistics Volume 6 ed. Beachwood, Ohio, Institute of Mathematical Statistics, 2000.

Web References

- EAGLET: www.columbia.edu/~ws2267/SOFT/soft.html
- CALEB: <http://onion.cpmc.columbia.edu/caleb/>
- MORGAN: www.stat.washington.edu/thompson/Genepi/pangaea.shtml