Multipoint and Single Point Non-Parametric Linkage Analysis With Imperfect Data

Patrick F. Sullivan,1* Benjamin M. Neale,2 Michael C. Neale,2 Edwin van den Oord,2 and Kenneth S. Kendler2
1Departments of Genetics and Psychiatry, University of North Carolina at Chapel Hill, North Carolina
2Virginia Institute for Psychiatric & Behavioral Genetics, Virginia Commonwealth University, Virginia

We used simulation to explore the impact of common data imperfections (i.e., missing parents, genotyping error, map error, and missing genotypes) upon the performance of multipoint and single point linkage analysis in the analyses of linkage data from pairs of siblings affected with an idealized complex trait. The performance of single point and multipoint linkage was similar under an unrealistic best case scenario; however, when four data imperfections were combined, the performance of single point linkage analysis appeared to be superior to multipoint. The absence of parental genotypes in the presence of 1% genotype error led to marked degradation of linkage signal, particularly for multipoint analyses. © 2003 Wiley-Liss, Inc.

KEY WORDS: linkage; multipoint; non-parametric; genotype error; simulation

INTRODUCTION

In the last decade, genetic linkage analysis using the genome scan approach has been an important discovery technique for complex traits of public health significance. For example, Altmüller et al. [2001] identified 101 genome scans for 31 human complex traits published between 1993–2000, half of which studied affected sibling pairs. Despite skepticism about its utility [Terwilliger and Göring, 2000], there appears to have been considerable recent progress in identifying polymorphisms that confer risk for a number of complex traits using linkage analysis [Korstanje and Paigen, 2002], including notably difficult traits like schizophrenia [Stefansson et al., 2002; Straub et al., 2002; van den Oord et al., in press].

Multipoint non-parametric (“model free”) linkage analysis is routinely used in genome scans for complex traits [Elston and Cordell, 2001]. Indeed, the development and distribution of software that could efficiently conduct multipoint linkage analyses was generally viewed as an important advance over traditional pairwise or “single point” analyses because of the more efficient use of marker and linkage information [Elston, 1992; Kruglyak and Lander, 1995; Weeks et al., 1995; Kruglyak et al., 1996]. However, initial reports that demonstrated the superiority of multipoint over single point non-parametric linkage analysis typically used idealized data sets lacking imperfections commonly found in practice. For example, modest levels of genotyping error can lead to considerable loss of power to detect linkage [Douglas et al., 2002], and multipoint non-parametric linkage analysis is sensitive to the presence of genotyping error [Göring and Terwilliger, 2000; Abecasis et al., 2001]. Multipoint analysis requires a reasonably accurate genetic map and map error can lead to bias and inefficiency [Buetow, 1991; Halpern and Whittemore, 1999; Daw et al., 2000].

The goal of this report was to compare the performance of multipoint and single point non-parametric linkage analyses in simulated data that contain different imperfections commonly found in linkage data sets of complex traits. A motivation for this work was the observation of lesser mean IBD sharing across the genome in multipoint (0.48) than in single point (0.50) non-parametric linkage in a genome scan for nicotine dependence [Straub et al., 1999] and by the inability of the literature completely to explain this observation.

METHODS

We used GASP [Wilson et al., 1996] to simulate pedigrees consisting of two genotyped parents and two
genotyped and phenotyped offspring. We simulated a 170 cM chromosome with markers every 10 cM (18 total markers). The simulated markers were all microsatellites with six alleles (heterozygosity 0.82, allele frequencies of 0.1, 0.1, 0.2, 0.2, 0.2, and 0.2). The genetic model was chosen as an optimistic approximation of a generic complex disorder of 10% prevalence. A biallelic quantitative trait locus was situated at 45 cM whose risk-increasing allele frequency was 0.25. The variance components model consisted of 50% additive genetic effects (the quantitative trait locus plus residual genetic effects) and 50% unshared environmental effects. GASP was used to simulate families until a specified number of affected sibling pairs (ASP) accumulated. ASPs were defined as two siblings who both had trait liability scores in the top 10% of the distribution. This model assumed additive effects, random mating, and the absence of selection. One thousand simulated samples were generated for each of four conditions: 200 ASPs with a 0% QTL, 100 ASPs/12.5% QTL, 200 ASPs/12.5% QTL, and 400 ASPs/12.5% QTL.

In the “best case” scenario, the data analyzed were those generated by GASP without additional imperfections. Next, the best case scenario samples were degraded in one of six ways prior to linkage analysis: (1) We modified the best case scenario by setting all parental genotypes to missing as might occur when studying sibling pairs affected for a disorder whose onset is late in life (e.g., Alzheimer disease); (2) The best case scenario was modified by the addition of 1% error per genotype. If a pseudorandom number (uniform, range: 0–1) ≤ 0.01 was generated, then true heterozygotes were converted into homozygotes (e.g., 4/5 became either 4/4 or 5/5 at random). Alternatively, homozygotes were shifted by one allele (e.g., 3/3 became either 2/2 or 4/4 at random). These two types of genotyping error are particularly worrisome and perhaps more probable when demands for high-throughput genotyping dictate the use of relatively constant genotyping conditions across markers. Conversion of true heterozygotes into homozygotes can occur when the second allelic band is faint or attributed to stutter and shifting of homozygotes by one band can occur via misreading of a banding pattern; (3) The best case scenario was modified by the inclusion of map error (markers whose true location was at 50 and 90 cM were inverted). This error will only influence multipoint analysis as the genetic model is generally unknown for complex traits and focused on LODmax as the linkage estimate of critical interest to researchers investigating complex traits. Allele frequencies were estimated from the data.

RESULTS

The simulation results for LODmax are depicted in Figure 1 which summarizes 56,000 MERLIN runs (4 conditions × 1,000 replicates × 7 modifications × 2 modes of linkage analysis). Each panel in Figure 1 compares six degradations from the best case scenario for the same 1,000 simulated samples. A pair of bars summarizes the distribution of LODmax scores (the vertical bars extend from the 2.5th–97.5th percentiles and the median is indicated by a horizontal tick mark) with single point linkage results on the left and multipoint on the right. We wish to highlight the following five features of Figure 1.

Main Analyses

First, we considered the impact of the degradations that were introduced. In comparison to the best case scenario, the addition of 1% genotyping error or 5% missing genotypes made little difference to the LODmax distribution. In contrast, omission of parental genotypes led to an appreciable downward shift in the LODmax scores. The impact of the addition of 1% genotyping error depended on whether parental genotypes were present. The introduction of 1% genotyping error as an isolated feature had minimal impact as MERLIN detected and removed ~80% of the errors when parents were genotyped; however, in the absence of parental genotypes, only ~44% of the genotyping errors were detected and the addition of 1% genotyping error led to a marked downward shift in the LODmax distribution. Finally, the introduction of map error (switching of markers at 50 and 90 cM with the trait locus at 45 cM) had a small effect (single point analyses are independent of the genetic map but are shown for comparison).

Second, the median multipoint LODmax was less than the single point median LODmax in 25 of the 28 pairs in Figure 1 (the three exceptions are in panel d and are of small magnitude). If there is no true genetic effect (panel a), this is advantageous as multipoint linkage may be less likely to yield a false positive finding; however, if there is a true genetic effect (panels b, c, and d) then multipoint linkage could possess less power to detect a true effect.

Third, as anticipated, LODmax declines sharply as the data set is degraded from best to worst case. The LODmax distributions are generally similar for multipoint and single point linkage in the best case scenarios, but, in the worst case scenarios, the declines in LODmax are twice as large for multipoint than for single point non-parametric linkage analysis.

Fourth, the variability in LODmax (indexed by the difference between the 2.5th and 97.5th percentiles) was
greater for single point than for multipoint analyses in the two most unfavorable scenarios in Panels a–d of Figure 1 (absence of parental genotypes with 1% genotyping error and the worst case scenario).

Fifth, the lower bound of the vertical bars in Figure 1 is an index of statistical power in Panels b–d of Figure 1. For single point linkage, the lower bounds of the LOD$_{\text{max}}$ distributions are always greater than the analogous multipoint linkage results.

### QTL Localization

We compared the capacity of single point and multipoint linkage analyses to identify the true location of the simulated QTL. Using the simulated data in Figure 1, we noted the position of LOD$_{\text{max}}$ and computed its mean and standard deviation over 1,000 simulated trials. To simplify the comparison, we analyzed only the best and worst case scenarios. Figure 2 depicts six comparison pairs (single point versus multipoint for best/worst case scenarios over the parameters from the Panels b–d in Figure 1). For the three pairs of best case scenarios corresponding to Panels b–d in Figure 1, multipoint linkage evidenced superior localization of the QTL (i.e., location means closer to the true location and smaller standard deviations). For the worst case scenarios, multipoint tended to localize the QTL better than for single point analyses but the standard deviation of the position of LOD$_{\text{max}}$ was considerably larger. As expected, the QTL was not well-localized in any of these analyses.

### Additional Simulations

(a) We examined dominant and recessive genetic models for 200 ASPs and a 12.5% QTL (data not shown). The pattern of results was similar to those for the additive case in Panel c of Figure 1. (b) We also examined the LOD$_{\text{mean}}$ distributions (the mean LOD over a simulated chromosome). In the absence of a true genetic effect, the expectation of the LOD$_{\text{mean}}$ is zero which was closely approximated in both multipoint and single point linkage results under the best case scenario. For the worst case scenario, LOD$_{\text{mean}}$ was shifted.
downward and the effect was greater for multipoint (LOD\textsubscript{mean} = -0.35) than for single point linkage analyses (LOD\textsubscript{mean} = -0.09). If a true genetic effect was present, the downward shift from best to worst case was always greater for multipoint than single point linkage results. (c) Examination of non-parametric Z\textsubscript{max} scores (the maximum NPL Z score in a simulated chromosome) revealed a similar pattern of results as

![Plot of the standard deviation versus mean for the location of LOD\textsubscript{max} computed over 1,000 simulated samples. The simulated QTL was located at 45 cM. There are six pairs of points corresponding to the best case/worst case scenarios for the numbers of ASPs and QTL effect size depicted in panels b–d of Figure 1. The black lines connect the data from the single point and multipoint analyses of the same set of simulated data.](image-url)
for the non-parametric LOD\textsubscript{max} scores (data not shown). For example, in the absence of a genetic effect in the worst case scenario, the median Z\textsubscript{max} was considerably lower for multipoint (−0.88) than for single point linkage (−0.40).

(d) To enhance the comparability of single point and multipoint analyses, we restricted multipoint estimation to the marker locations. In practice, multipoint is often used to estimate linkage statistics at several inter-marker locations and this could enhance the performance of the multipoint approach. To evaluate this possibility, we ran additional multipoint simulations for the best and worst case scenarios in panel c of Figure 1 (200 ASPs, 12.5% QTL effect, 1000 simulations for each) but with linkage statistics estimated every 2 cM instead of every 10 cM. This change had little effect: for the best case scenario, the LOD\textsubscript{max} results were 0.15/1.07/3.06 (2.5th centile/median/97.5th centile) for multipoint estimation every 2 cM vs. 0.15/1.04/2.86 for estimation every 10 cM. For the worst case scenario, the LOD\textsubscript{max} results were −0.02/0.23/1.40 for multipoint estimation every 2 cM vs. −0.02/0.23/1.38 for estimation every 10 cM.

**DISCUSSION**

The goal of these simulations was to compare the performance of multipoint and single point non-parametric linkage analysis in the presence of a number of data imperfections commonly found in practice. In comparison to the best case scenario, the impact of data imperfections in isolation—e.g., absence of parental genotypes, 1% genotyping error, map error, and 5% missing genotypes—did not have a strong impact on the performance of either multipoint or single point linkage analysis. However, when these degradations were present in combination, the impact on LOD\textsubscript{max} was clearly more pronounced for multipoint than single point non-parametric linkage. As in a prior report, the absence of parental genotypes combined with 1% genotype error was a particularly potent source of bias and loss of information [Abecasis et al., 2001].

For samples consisting of ASPs without parental genotypes, the results in Figure 1 imply that multipoint non-parametric linkage is a double-edged tool. Multipoint analysis may act as a smoothing technique [Teng 2002] but with linkage statistics estimated every 2 cM instead of every 10 cM. This change had little effect: for the best case scenario, the LOD\textsubscript{max} results were 0.15/1.07/3.06 (2.5th centile/median/97.5th centile) for multipoint estimation every 2 cM vs. 0.15/1.04/2.86 for estimation every 10 cM. For the worst case scenario, the LOD\textsubscript{max} results were −0.02/0.23/1.40 for multipoint estimation every 2 cM vs. −0.02/0.23/1.38 for estimation every 10 cM.

As shown in Figure 2, multipoint analyses tended to localize the QTL better than single point analyses. However, under the worst case scenario, this advantage was accompanied by the important disadvantage of an inflated standard deviation of the estimate of the location of the QTL. Thus, for localization, multipoint non-parametric linkage yielded mixed results.

We chose not to perform a highly exhaustive search of the parameter space possible for models of complex traits as we wished to demonstrate concisely that both single point and multipoint non-parametric linkage analyses should be routinely conducted. It is possible that a more exhaustive search of the parameter space could yield different conclusions. However, the genetic models we analyzed are plausible for many complex traits: given that the underlying genetic model is usually unknown, we suggest that our results raise a valid analytic concern of general importance. It is possible that these conclusions might not apply to multiple generation pedigrees or affected sibships larger than size two; however, ASPs predominate in many linkage samples for complex traits [Altmüller et al., 2001].

Finally, we note that the extraction of nearly as much information from single point as multipoint analyses had been anticipated [Morton et al., 1988]. Although made in a different context, the recommendation that pairwise LOD scores should be examined before multipoint LOD scores remains appropriate [Risch and Giuffra, 1992].

**REFERENCES**


94 Sullivan et al.
