Approximation of Multipoint Likelihoods Using Flanking Marker Data: A Simulation Study

Andrew W George[1], LaVonne A. Mangin[2], Christopher W. Bartlett[1,3], Mark W. Logue[1], Alberto M. Segre[1,2], and Veronica J. Vieland[1,4]

1. Program in Public Health Genetics, College of Public Health

2. Department of Computer Science, College of Liberal Arts and Sciences

3. Department of Internal Medicine, Roy L. and Lucille A. Carver College of Medicine

4. Department of Psychiatry, Roy L. and Lucille A. Carver College of Medicine

University of Iowa, Iowa City IA 52242

Email Addresses

Andrew George: andrew-george@uiowa.edu
LaVonne Mangin: lavonne_mangin@msn.com
Christopher Bartlett: christopher-bartlett@uiowa.edu
Mark Logue: mark-logue@uiowa.edu
Alberto Segre: alberto-segre@uiowa.edu
Veronica Vieland: veronica-vieland@uiowa.edu

Abstract

The calculation of multipoint likelihoods is computationally challenging with the exact calculation of multipoint probabilities only possible on small pedigrees and many markers or large pedigrees and few markers. This paper explores the utility of approximating multipoint likelihoods using data on markers flanking a hypothesized position of the trait locus. The calculation of such likelihoods is often feasible, even on large pedigrees with missing data and complex structures. Performance characteristics of the approximation procedure are assessed through the calculation of multipoint Heterogeneity Lod scores on data simulated for Genetic Analysis Workshop 14. Analysis is restricted to data on the Aipotu population on chromosomes 1, 3 and 4 where chromosomes 1 and 3 are known to contain disease loci. The approximation procedure performs well, even when missing data and genotyping errors are introduced.

Introduction

The calculation of multipoint likelihoods on general pedigrees is computationally challenging. Factors influencing the complexity of multipoint calculations include family size, pedigree structure, marker number and missing data. Efficient algorithms have been developed for handling large pedigrees and few markers [1] or small pedigrees and many markers [2] but calculating multipoint probabilities on large pedigrees and many markers is infeasible.

In this paper, the performance characteristics of FLANK, a computer program for approximating multipoint likelihoods on general pedigrees and many linked markers, is explored. Multipoint likelihoods are approximated using data observed only on markers flanking a hypothesized position of the trait locus. Calculating these three-point likelihoods is often feasible even on large complex pedigrees. Likelihood computations in FLANK are performed via VITESSE [3]. The speed and accuracy of FLANK are examined through a Heterogeneity Lod (Hlod) score analysis of data simulated on the Aipotu population from Genetic Analysis Workshop (GAW) 14.

Methods

FLANK

FLANK is a computer program, using VITESSE to perform the likelihood calculations, to approximate multipoint likelihoods on general pedigrees. Likelihoods are approximated from multipoint calculations only on marker data flanking a hypothesized position of the trait locus. Results are reported as Homogeneity Lod scores or Heterogeneity Lod Scores (Hlod) at pre-specified positions of the trait locus. Functionally, FLANK is similar to GENEHUNTER [4]. Only a single locus file and pedigree file in linkage format are required as input files. Command line arguments are used to specify the distance (in cM) between hypothesized positions, the distance beyond the linkage map (if any) to compute likelihoods, and whether Homogeneity Lod scores or Hlod scores are to be reported.

Simulated Data

In this study, data generated on the Aipotu population for GAW14 are chosen for analysis. The Aipotu population consists of 100 nuclear families, ranging in size from 2 to 10 siblings. Marker and trait data are observed on each individual. The trait is dichotomous where an individual is either affected or unaffected for the disease. Microsatellite marker data on chromosomes 1, 3 and 4, containing 41, 42, and 44 linked markers respectively, are selected for analysis. Inter-marker distances, on average, are 7.5cM. Chromosome 1 contains a disease locus between the 23rd and 24th marker locus. Chromosome 3 contains a disease locus between the 41st and 42nd marker locus and Chromosome 4 is unlinked to disease causing loci. There are 100 replicates of the data.

Linkage Detection and Mapping

The accuracy of FLANK for detecting and localizing trait loci is examined through the analysis of simulated family data. A dominant trait model with incomplete penetrance and a disease allele frequency of 0.01 is assumed. Here, three marker sets formed from the original GAW14 simulated data are considered: four linked markers closest to the disease locus (Mset4), 16 linked markers closest to the disease locus (Mset16), and all markers on a chromosome (MsetAll). Since Chromosome 4 is unlinked to a disease locus, markers are selected from the beginning of the marker map. Hlod scores are calculated every 1cM. The accuracy of the approximated Hlod scores is determined through comparison with exact multipoint Hlod scores calculated using GENEHUNTER.

Missing Data and Genotyping Errors

Multipoint calculations on pedigrees are affected by missing data and genotyping error. To explore the accuracy of FLANK given imperfect data, missing data and Mendelian consistent genotyping errors are introduced. The marker phenotype at a locus for an individual is randomly removed with probability 0.01. Mendelian consistent genotyping errors are created, with probability 0.005, by randomly permuting with equal probability the transmitted allele from one of the parents. Note that this error model is simplistic since it cannot produce genotyping errors in the parents and does not make distinctions between types of genotyping errors, which are all equally likely in the present study. The assumed probability of Mendelian consistent errors is consistent with an overall (pedigree consistent and inconsistent) genotyping error rate of 1% [5]. The levels of missing data and genotyping error are realistic compared to real datasets [6].

Results

Linkage Detection and Mapping

To examine the accuracy of approximating multipoint likelihoods with flanking data, mean Hlod scores averaged over replicates are calculated at each hypothesized position of the trait locus for chromosomes 1, 3 and 4 for marker sets Mset4, Mset16 and MsetAll. There is close agreement between the exact and approximated Hlod scores across hypothesized positions of the trait locus. The mean scores at the known location of the trait locus for chromosomes 1 and 3 and at an arbitrary chromosomal position for Chromosome 4 are reported in Table 1. Given the associated standard errors, the difference between associated scores is insignificant. Also, there is little change in mean exact multipoint Hlod scores across marker sets.

Table 1. Comparison of exact and approximate Hlod scores for different marker sets.
Mean exact Hlod scores ($Hlod_{mult}$) and mean approximate Hlod ($Hlod_{flnk}$) scores, averaged over 100 simulated replicates of the Aipotu population, at a location between markers flanking the trait locus for chromosomes 1 and 3. For Chromosome 4, mean Hlod scores are reported at an arbitrary chromosomal position, common to all three marker sets Mset4, Mset16 and MsetAll. Exact multipoint scores are calculated on all available markers jointly. Approximate multipoint scores are calculated only on flanking markers jointly. Standard errors of the means are given in parentheses.

		Mset4		Mset16		MsetAll	
Chrm	Chrml Pos	$\mathrm{Hlod}_{\mathrm{flnk}}$	$\mathrm{Hlod}_{\mathrm{mult}}$	$\mathrm{Hlod}_{\mathrm{flnk}}$	$\mathrm{Hlod}_{\mathrm{mult}}$	$\mathrm{Hlod}_{\mathrm{flnk}}$	$\mathrm{Hlod}_{\mathrm{mult}}$
1	175cM	1.96 (0.15)	2.23 (0.15)	1.95 (0.15)	2.30 (0.16)	1.94 (0.15)	2.30 (0.16)
3	312cM	1.57 (0.14)	1.57 (0.14)	1.57 (0.14)	1.57 (0.14)	1.57 (0.14)	1.57 (0.14)
4	20cM	0.06 (0.02)	0.06 (0.02)	0.06 (0.02)	0.06 (0.02)	0.06 (0.02)	0.06 (0.02)

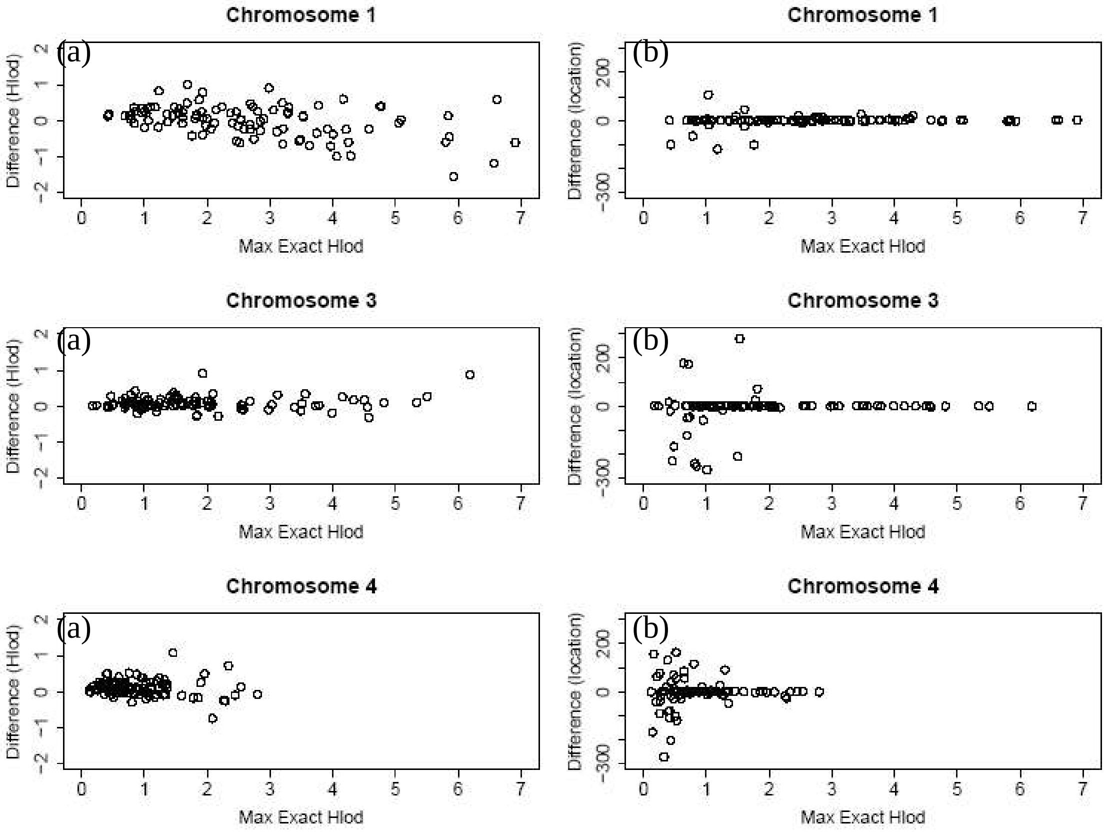
To examine the utility of FLANK for detecting and localizing trait loci, differences in the peak exact and approximated Hlod scores and differences in the chromosomal location of the peaks are investigated. Figure 1a plots the difference in peak Hlod scores on the vertical axis against the peak exact multipoint Hlod on the horizontal axis for the analysis of data on MsetAll. That is, each point represents the difference in approximate and exact linkage results from the analysis of a single replicate. In Figure 1a, there are a cluster of points about a horizontal line intersecting 0 on the vertical axis indicating near perfect multipoint estimates. Points off the horizontal line indicate differences in peak scores. However, it is reassuring that the largest differences in peak scores occur when the peak exact score is also large. Using Hlod scores calculated on flanking markers for detection does not result in conclusions that are different to analyzing data on all available markers jointly.

Figure 1b plots the difference in the chromosomal location of the peaks on the vertical axis against the peak exact multipoint Hlod on the horizontal axis. Again, there is a clustering of points about a horizontal line intersecting 0 on the vertical axis indicating close agreement between the localization of the trait using flanking markers and all available markers. It is also reassuring that the largest differences in locations occur for small peak exact scores. When the peak exact score is small, there is little information in the data for estimating linkage. Using Hlod scores calculated on flanking markers for localization does not result in conclusions that are different to analyzing data on all available markers jointly.

Missing Data and Genotyping Errors

Results are reported for the analysis of Chromosome 1 where missing data and genotyping errors are randomly introduced. Results from the analysis of Chromosome 3 and 4 data are not reported but are consistent with the Chromosome 1 analysis. Mean Hlod scores at hypothesized positions of the trait locus for data on Chromosome 1 for the three markers sets are plotted in Figure2. Each plot contains the mean exact multipoint Hlod scores on data without error or unobserved information (the circles) and a single mean Hlod score curve (the dashed line). The mean Hlod score curve represents the analysis of data with both 1% missing data and 0.5% genotyping error. This curve is indistinguishable from the mean Hlod score curves on data with only 1% missing data or 0.5% genotyping error (not shown). Here, the missing data and/or genotyping error has no discernable affect on linkage results. Furthermore, for hypothesized positions of the trait locus common to the three marker sets, there is little change in mean exact Hlod scores.

Figure 1. **Differences in peak Hlod scores and differences in inferred locations of trait locus.**
(a) Difference between the exact peak Hlod score and the approximate peak Hlod score against the exact peak Hlod score for analyses of data on chromosomes 1, 3 and 4. (b) Difference between the chromosomal locations of the peak Hlod scores against the exact peak Hlod score for analyses of data on chromosomes 1, 3 and 4. Each point represents results from the analysis of a single data replicate for marker set MsetAll.

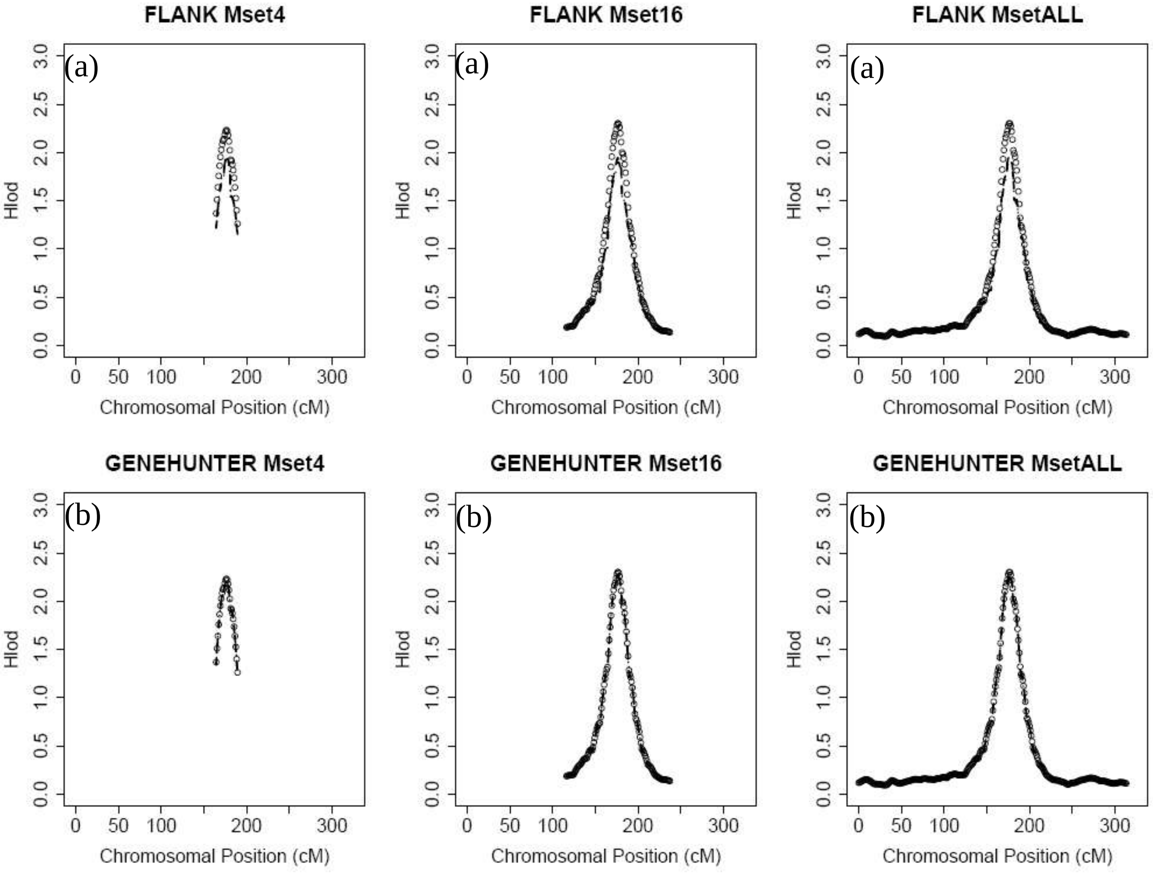


Conclusion

In this paper, approximating multipoint likelihoods using a new computer program FLANK is accessed through the calculation of Hlod scores on simulated data. By only considering data observed on flanking markers, the computational complexity of multipoint calculations are greatly reduced. For data simulated on nuclear families, there is little loss in accuracy using the proposed approximation procedure. Furthermore, FLANK produced multipoint estimates, on average, an order of magnitude faster than GENEHUNTER. Further exploration is warranted on extended families, differing amounts and patterns of missing data, differing amounts of genotyping error and changes in marker informativeness.

Figure 2. Comparison of mean Hlod scores calculated using FLANK and GENEHUNTER on Chromosome 1 data

(a) Mean approximated Hlod scores, calculated using FLANK, on Chromosome 1 data. (b) Mean exact Hlod scores, calculated using GENEHUNTER, on Chromosome 1 data. Circles represent exact multipoint Hlod scores, calculated on Chromosome 1 data without errors or missing data. The dashed line represents the Hlod score curve calculated on data with 1% missing data and 0.5% Mendelian consistent genotyping error.



References

[1] Elston R, Stewart J: A general model for the analysis of pedigree data. *Hum Hered* 1971, **21:** 523-542.

[2] Lander E, Green P: Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.* 1987, **84:** 2363-2367.

[3] O'Connell J, Weeks D: The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet.* 1995, **11:** 402-408.

[4] Kruglyak, L, Daly M, Reeve-Daly M, Lander E: Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am. J. Hum. Genet.* 1996, **58:** 1347-1363.

[5] Douglas, J, Skol A, Boehnke M: Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet.* 2002, **70:** 487-95.

[6] Brzustowicz L, Merette C, Xie X, Townsend L, Gilliam T, Ott J: **Molecular and statistical approaches to the detection and correction of errors in genotype databases**. *Am. J. Hum. Genet*.1993, **53:** 1137-45.