A survey of current software for linkage analysis

Frank Dudbridge

MRC Human Genome Mapping Project Resource Centre, Hinxton, Cambridge CB10 1SB, UK; Tel: +44 1223 494572;
Fax: +44 1223 494512; E-mail: f.dudbridge@hgmp.mrc.ac.uk

Date received (in revised form): 18th August 2003

Abstract

There is now a wide choice of software available for linkage analysis. The most well known packages are briefly reviewed here. The package with the most extensive range of analyses is GENEHUNTER, but for many of its functions there are other programs with better performance. These include FASTLINK and VITESSE for parametric analysis ALLEGRO and MERLIN for non-parametric analysis and SOLAR for variance components analysis. The computational limits of current approaches can be improved with SIMWALK2 and the promising new SUPERLINK program. Directions for future work include improved user interfaces and consensus formats for data input and exchange.

Keywords: *software, linkage analysis, programming*

There is now a wide choice of methods and software available for mapping genes by linkage. Although the method of analysis is often determined by the experimental design, there is less guidance regarding the most appropriate software. Here, the most well-known packages for linkage analysis will be briefly reviewed and some directions and standards for future work will be suggested.

At one extreme, linkage analysis is applied to a small number of large pedigrees in which the trait exhibits a strongly Mendelian mode of inheritance. Methods for this type of data are usually termed 'parametric' because an explicit penetrance model defining the relationship between genotype and disease must be specified. The most flexible package for these analytical methods remains FASTLINK,[1,2] which is functionally equivalent to the original LINKAGE package.[3] For most pedigree structures, whether one applies single- or multi-point analysis of a disease or quantitative trait, VITESSE is a faster package;[4,5] however, FASTLINK continues to be more efficient for pedigrees containing inbreeding loops.

At the other extreme, linkage analysis is also applied to a large number of small pedigrees with unknown mode of inheritance. 'Non-parametric' allele-sharing methods are usually preferred here, for which the most well-known program is GENEHUNTER.[6,7] GENEHUNTER contains an extensive set of linkage and association tests and, as such, is a *de facto* standard for statistical genetics analysis.[8] A disadvantage of this position is that any new program will aspire to improve on GENEHUNTER, so that for many of its functions there are now other programs with better performance. An important example is ALLEGRO,[9] which is faster for most pedigree structures, includes a wider range of scoring functions and computes more accurate significance levels for non-parametric statistics. The latter feature is also available in GENEHUNTER-PLUS,[10] but this is only available for version 1.3 of GENEHUNTER and so does not access the speed-ups available in later versions.

Another recent competitor is MERLIN,[11] which employs a still faster algorithm that is particularly useful in dense marker maps, for which the number of recombinations allowed between markers can be constrained. The range of analyses is similar to GENEHUNTER, MERLIN also provides the linear-model lod score available in ALLEGRO but not the exponential model. MERLIN does not calculate parametric lod scores — which are available in GENEHUNTER and ALLEGRO — but for non-parametric analysis, error checking and haplotyping, it will often be the fastest program. All three of these programs handle X-linked data, although this also is only available in version 1.3 of GENEHUNTER.

An alternative approach for an unknown mode of inheritance is to perform parametric analysis over a range of models and then adjust the best lod score for this optimisation. This approach is implemented in MFLINK.[12] In small pedigrees, there seems to be little to choose between this approach and the allele-sharing methods discussed above;[13] however, currently MFLINK can only perform two-point analysis.

A promising new model is implemented in SUPERLINK.[14] Fishelson and Geiger show that the algorithms used by FASTLINK and GENEHUNTER are instances of a more general model, under which a more efficient order of computation is determined at run-time according to the input pedigree. For parametric linkage analysis, some impressive speed-ups over VITESSE have been reported. Future versions will include allele sharing and other statistics (M. Fishelson, personal communication).

Quantitative traits are commonly analysed by regression or by variance-components methods. Haseman–Elston regression is a sib-pair method available in GENEHUNTER with heuristic adjustments for general pedigrees. Recently, the regression framework has been extended to more general pedigrees,[15] and this is implemented in MERLIN. This approach now has comparable power to variance-components methods, with less dependence on trait normality and some computational advantages. MERLIN and GENEHUNTER also provide rank-based tests (confusingly also termed 'non-parametric'), which are appropriate for non-normally distributed traits. Again, note that for GENEHUNTER the test is a sib-pair method, with heuristic adjustments for general pedigrees, whereas for MERLIN the test is immediately applicable to general pedigrees.

Variance-components methods are more powerful than regression, provide parameter estimates and easily accommodate a wide range of null hypotheses; the cost is stronger dependence on trait normality and higher computational burden. Implementations are available in MERLIN, provided that no dominance variance is assumed, and in GENEHUNTER. Another very flexible package for variance components model fitting is SOLAR.[16] MERLIN is currently the only program that can perform multipoint variance components analysis on the X chromosome. ALLEGRO also contains undocumented implementations of various quantitative trait methods.

Exact multipoint analysis is limited either by the number of markers that can be included (FASTLINK, VITESSE) or the pedigree size (GENEHUNTER, ALLEGRO, MERLIN). With current microsatellite markers, large pedigrees usually contain enough information from a small number of markers for current software to be adequate. This will change with the move to automated single nucleotide polymorphism typing for linkage studies,[17] so it is becoming more important to have software that can handle large numbers of markers in large pedigrees. Currently, this is only generally possible through the approximation methods of SIMWALK2, which nevertheless has good reported accuracy.[18] Although the program has a lot of tuning parameters, the MEGA2 utility program provides a reasonably easy route to a default analysis which is suitable in most cases.[19] More efficient approximation methods are an area of current research, for example MORGAN,[20] which currently only allows fully penetrant recessive traits but shows promise for more general models.

Modern computing favours graphical user interfaces (GUIs), which allow mouse-driven input; but these are conspicuously absent from linkage software. Descendents of LINKAGE have essentially no user interface, although the terminal-based tool LCP is available to set up analysis scripts; GENEHUNTER and SOLAR run their own interactive command shells, whereas ALLEGRO and MERLIN use a single command with optional arguments and auxiliary input files. On the plus side, all of these interfaces are amenable to scripting — for example to allow one to repeat the same analysis on multiple input files — but the single-command interface of ALLEGRO and MERLIN is easily the most convenient to use in scripts. With the availability of Java, HTML and TCL as cross-platform languages for GUI development, it is hoped that future versions of these packages will incorporate simpler user interfaces, as well as scriptable back ends.

The LINKAGE input file format is recognised by many programs but is by no means universal. MEGA2 is a useful utility for converting between formats, but even this requires an additional map file which duplicates information contained in the locus file. It is hoped that the LINKAGE format, however imperfect, will eventually be recognised by all programs that perform linkage analysis, without the need for supplementary conversion scripts.

GENEHUNTER, ALLEGRO, MERLIN and SOLAR can all output multipoint identical-by-descent (IBD) distributions, which are valuable for gaining insights into the segregation patterns in pedigrees. None can input this information, however: it is not possible, say, to calculate the IBD distribution under the recombination restrictions of MERLIN and then use this to obtain an exponential-model lod score from ALLEGRO. Furthermore, sometimes different analyses result in the same distribution, and it is inefficient to recompute it each time. With some caveats, it is possible to avoid this recomputation in SOLAR, but simple input of IBD, haplotype and recombination information would still generally be a useful feature for future versions.

This survey has necessarily been cursory, and there is a wealth of other good linkage software available. Two internet sites provide useful lists of available software. A comprehensive list of statistical genetics software can be found at http://www.nslij-genetics.org/soft/, with links to their sources. This list continues to be mirrored at its previous site, http://linkage.rockefeller.org/soft/. It is perhaps over-inclusive, containing a number of obsolete programs, and it makes no recommendations. By contrast, the collection at http://www.hgmp.mrc.ac.uk/Registered/Menu/linkage.html contains only the most popular programs, but provides executable files, browsable documentation and a web-based graphical interface for the most common applications.

References

1. Cottingham, R.W., Idury, R.M. and Schäffer, A.A. (1993), 'Faster sequential genetic linkage computations', *Am. J. Hum.Genet.* Vol. 53, pp. 252–263.
2. Schäffer, A.A., Gupta, S.K., Shiram, K. and Cottingham, R.W. (1994), 'Avoiding recomputation in linkage analysis', *Hum. Hered.* Vol. 44, pp. 225–237.
3. Lathrop, G.M. and Lalouel, J.M. (1984), 'Easy calculations of lod scores and genetic risks on small computers', *Am. J. Hum. Genet.* Vol. 36, pp. 460–465.
4. O'Connell, J. R. and Weeks, D. E. (1995), 'The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype and set-recoding and fuzzy inheritance', *Nat. Genet.* Vol. 11, pp. 402–408.
5. O'Connell, J.R. (2001), 'Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm', *Hum. Hered.* Vol. 51, pp. 226–240.
6. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996), 'Parametric and non-parametric linkage analysis: A unified multipoint approach', *Am. J. Hum. Genet.* Vol. 58, pp. 1347–1363.
7. Markianos, K., Daly, M.J. and Kruglyak, L. (2001), 'Efficient multipoint linkage analysis through reduction of inheritance space', *Am. J. Hum. Genet.* Vol. 68, pp. 963–977.
8. Nyholt, D.R. (2002), 'GENEHUNTER: Your 'one-stop shop' for statistical genetic analysis?', *Hum. Hered.* Vol. 53, pp. 2–7.
9. Gudbjartsson, D.F., Jonasson, K., Frigge, M. and Kong, A. (2000), 'Allegro, a new computer program for multipoint linkage analysis', *Nat. Genet.* Vol. 25, pp. 12–13.
10. Kong, A. and Cox, N.J. (1997), 'Allele-sharing models: LOD scores and accurate linkage tests', *Am. J. Hum. Genet.* Vol. 61, pp. 1179–1188.
11. Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002), 'Merlin – rapid analysis of dense genetic maps using sparse gene flow trees', *Nat. Genet.* Vol. 30, pp. 97–101.
12. Curtis, D. and Sham, P.C. (1995), 'Model-free linkage analysis using likelihoods', *Am. J. Hum. Genet.* Vol. 57, pp. 703–716.
13. Sham, P.C., Lin, M.W., Zhao, J.H. and Curtis, D. (2000), 'Power comparison of parametric and non-parametric linkage tests in small pedigrees', *Am. J. Hum. Genet.* Vol. 66, pp. 1661–1668.
14. Fishelson, M. and Geiger, D. (2002), 'Exact genetic linkage computations for general pedigrees', *Bioinformatics* Vol. 18(Suppl. 1), pp. S189–S198.
15. Sham, P.C., Purcell, S., Cherny, S.S. and Abecasis, G.R. (2002), 'Powerful regression-based quantitative-trait linkage analysis of general pedigrees', *Am. J. Hum. Genet.* Vol. 71, pp. 238–253.
16. Almasy, L. and Blangero, J. (1998), 'Multipoint quantitative trait linkage analysis in general pedigrees', *Am. J. Hum. Genet.* Vol. 62, pp. 1198–1211.
17. Matise, T.C., Sachidanandam, R., Clark, A.G. *et al.* (2003), 'A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set', *Am. J. Hum. Genet.* Vol. 73, pp. 271–284.
18. Sobel, E. and Lange, K. (1996), 'Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics', *Am. J. Hum. Genet.* Vol. 58, pp. 1323–1337.
19. Mukhopadyay, N., Almasy, L., Schroeder, M., Mulvihill, W. P. and Weeks, D. E. (1999), 'Mega2, a data-handling program for facilitating genetic linkage and association analyses', *Am. J. Hum. Genet.* Vol. 65(Suppl), p. A436.
20. George, A.W., Wijsman, E.M. and Thompson, E.A. (2002), 'Detecting disease genes via a new Markov chain Monte Carlo approach for multipoint linkage analysis', *Genet. Epidemiol.* Vol. 23, p. 283.