This is a preprint of a review article published in Human Genomics 1:63-65. The published article is avaiable at http://www.ingenta.com/journals/browse/hsp/hg/

A SURVEY OF CURRENT SOFTWARE FOR LINKAGE ANALYSIS

Frank Dudbridge

MRC Human Genome Mapping Project Resource Centre, Hinxton, Cambridge CB10 1SB

Tel: +44 1223 494572
Fax: +44 1223 494512
Email: f.dudbridge@hgmp.mrc.ac.uk

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 experiment design, there is less guidance regarding the most appropriate software. Here I shall briefly review the most well-known packages for linkage analysis, and suggest some directions and standards for future work.

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 [Cottingham et al., 1993; Schäffer et al., 1994] which is functionally equivalent to the original LINKAGE package [Lathrop et al., 1984]. For most pedigree structures, whether one applies single- or multi-point analysis of a disease or quantitative trait, VITESSE is a faster package [O'Connell and Weeks, 1995; O'Connell, 2001]; 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. "Nonparametric" allele-sharing methods are usually preferred here, for which the most well-known program is GENEHUNTER [Kruglyak et al., 1996; Markianos et al., 2001]. GENEHUNTER contains an extensive set of linkage and association tests, and as such is a *de facto* standard for statistical genetics analysis [Nyholt, 2002]. 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 other programs with better performance. An important example is ALLEGRO [Gudbjartsson et al., 2000], which is faster for most pedigree structures, includes a wider range of scoring functions, and computes more accurate significance levels for nonparametric statistics. The latter feature is also available in GENEHUNTER-PLUS [Kong and Cox, 1997], but this is only available for version 1.3 of GENEHUNTER and so does not access the speedups available in later versions.

Another recent competitor is MERLIN [Abecasis et al., 2002], 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, additionally providing the linear-model lodscore available in ALLEGRO but not the exponential model. MERLIN does not calculate parametric lodscores, which are available in GENEHUNTER and ALLEGRO; but for nonparametric 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 unknown mode of inheritance is to perform parametric analysis over a range of models, and then adjust the best lodscore for this optimization. This approach is implemented in MFLINK [Curtis and Sham, 1995]. In small pedigrees, there seems to be little to choose between this approach and

the allele-sharing methods discussed above [Sham et al., 2000]. However, currently MFLINK can only perform two-point analysis.

A promising new model is implemented in SUPERLINK [Fishelson and Geiger, 2002]. These authors 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 speedups over VITESSE have been reported. Future versions will include allele-sharing and other statistics [M. Fishelson, pers. comm.].

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 [Sham et al., 2002] 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 "nonparametric"), 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 no dominance variance is assumed, and in GENEHUNTER. Another very flexible package for variance components model fitting is SOLAR [Almasy and Blangero, 1998]. 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 SNP typing for linkage studies [Matise et al., 2003], so it is becoming more important to have software which 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 [Sobel and Lange, 1996]. 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 [Mukhopadyay et al., 1999]. More efficient approximation methods are an area of current research, for example MORGAN [George et al., 2002] which currently only allows fully penetrant recessive traits but shows promise for more general models.

Modern computing favours graphical user interfaces (GUI) which allow mouse-driven input; but these are conspicuously absent from the 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 recognized 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 recognized by all programs which 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. However none can input this information: 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 lodscore 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 the available software. A comprehensive list of statistical genetics software is 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. In 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

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97-101.

Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211.

Cottingham RW, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252-263.

Curtis D, Sham PC (1995) Model-free linkage analysis using likelihoods. Am J Hum Genet 57:703-716.

Fishelson M, Geiger D (2002) Exact genetic linkage computations for general pedigrees. Bioinformatics 18 Suppl 1:S189-S198.

George AW, Wijsman EM, Thompson EA (2002) Detecting disease genes via a new Markov chain Monte Carlo approach for multipoint linkage analysis. Genet Epidemiol 23:283.

Gudbjartsson DF, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12-13.

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179-1188.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and non-parametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363.

Lathrop GM, Lalouel JM (1984) Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460-465.

Markianos K, Daly MJ, Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 68:963-977.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, et al. (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet 73:271-284.

Mukhopadyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE. Mega2, a data-handling program for facilitating genetic linkage and association analyses. Am J Hum Genet 65 Suppl:A436.

Nyholt DR (2002) GENEHUNTER: your "one-stop shop" for statistical genetic analysis? Hum Hered 53:2-7.

O'Connell JR (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 51:226-240.

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype and set-recoding and fuzzy inheritance. Nat Genet 11:402-408.

Schäffer AA, Gupta SK, Shiram K, Cottingham RW (1994) Avoiding recomputation in linkage analysis. Hum Hered 44:225-237.

Sham PC, Lin MW, Zhao JH, Curtis D (2000) Power comparison of parametric and non-parametric linkage tests in small pedigrees. Am J Hum Genet 66:1661-1668.

Sham PC, Purcell S, Cherny SS, Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238-253.

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323-1337.