Invited Review

Human Heredity

Hum Hered 2002;53:2–7

Received: July 23, 2001
Revision received: September 6, 2001
Accepted: October 11, 2001

GENEHUNTER: Your 'One-Stop Shop' for Statistical Genetic Analysis?

Dale R. Nyholt

Queensland Institute of Medical Research, Brisbane, Australia

Key Words

GENEHUNTER · Linkage · Linkage disequilibrium · Model-based linkage analysis · Model-free multipoint linkage analysis · Variance components analysis · Transmission disequilibrium test · Quantitative trait loci

Abstract

The past decade has brought a proliferation of statistical genetic (linkage) analysis techniques, incorporating new methodology and/or improvement of existing methodology in gene mapping, specifically targeted towards the localization of genes underlying complex disorders. Most of these techniques have been implemented in user-friendly programs and made freely available to the genetics community. Although certain packages may be more 'popular' than others, a common question asked by genetic researchers is 'which program is best for me?'. To help researchers answer this question, the following software review aims to summarize the main advantages and disadvantages of the popular GENEHUNTER package.

Copyright © 2002 S. Karger AG, Basel

Traditionally, model-based linkage analysis, commonly known as logarithm of odds (lod) score analysis, was performed on multigenerational (extended) families segregating an obviously Mendelian (single-gene) trait. Lod score analysis is based on the likelihood ratio, which is the ratio between the probabilities of two alternatives L_{HA}/L_{H0}, where L_{H0} is the likelihood under the null hypothesis of no linkage (recombination fraction, $\Theta = 0.5$) and L_{HA} the likelihood of the alternative hypothesis of linkage ($\Theta < 0.5$). For a given set of pedigree genotype data, the likelihood of the observed data occurring is calculated, given a set of assumptions about the parameters of the underlying genetic model. These parameters include the allele frequencies of the loci involved, the relevant penetrance values and Θ between loci. Taking the base 10 logarithm ($\log_{10}$) of L_{HA}/L_{H0} produces the familiar lod score [1]. The maximum lod score is thus obtained by testing across different values of Θ.

The first such program written for use by genetic researchers was Ott's [2, 3] LIPED program. LIPED satisfactorily performs two-point (marker disease) lod score analysis. However, with the large number of markers now available for use in genetic linkage studies, it has become

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2002 S. Karger AG, Basel
0001–5652/02/0531–0002$18.50/0

Accessible online at:
www.karger.com/journals/hhe

Dr. Dale Nyholt
Queensland Institute of Medical Research
Post Office Royal Brisbane Hospital
Brisbane QLD 4029 (Australia)
Tel. +61 7 33620258, Fax +61 7 33620101, E-Mail daleN@qimr.edu.au

desirable to perform multipoint linkage analyses which utilize summed haplotype information from linked markers, thus increasing the amount of information regarding the cosegregation of a disorder with these markers.

In 1984, the LINKAGE set of programs [4–6] for estimation of recombination rates, calculation of lod score tables and analysis of genetic risks was released. The initial DOS version allowed model-based linkage analysis of an arbitrary number of loci, handling up to 20 alleles. Later releases allow the user to easily increase the number of alleles and other parameters after user-friendly recompilation. The packages' programs can (1) calculate two-point lod scores and risk with two or more loci (MLINK), (2) perform maximum likelihood estimation of the recombination rate and the maximum lod score from two-locus data (LODSCORE), (3) perform maximum likelihood estimation of recombination fractions, penetrance, gene frequencies and other parameters (ILINK), and (4) calculate multipoint location scores of one locus against a fixed map of other loci (LINKMAP).

In early 1993, a group of researchers converted the LINKAGE program code from 'Pascal' to the computationally faster 'C' programming language. This 'new' package was dubbed FASTLINK [7–9]. The FASTLINK package is essentially the same as LINKAGE, but roughly one order of magnitude faster on long runs. Input file preparation, program selection and analysis using the LINKAGE or FASTLINK programs are essentially the same.

When large numbers of multiallelic markers are analyzed in a pedigree, although the maximum possible number of haplotypes (MAXHAP) parameter may be increased and the program recompiled, multipoint linkage calculations are computationally complex and program limitations of the LINKAGE/FASTLINK programs are quickly reached. Alternatively, the VITESSE (meaning 'speed' in French) software package [10] can be used. VITESSE computes likelihoods with the functionality of the MLINK (two-point analysis) and LINKMAP (multipoint analysis) programs, but uses novel ('set-recoding' and 'fuzzy inheritance') algorithms to reduce the number of genotypes needed for exact computation of the likelihood, which accelerates the calculation. It also represents multilocus genotypes locus by locus to reduce the memory requirements. Although VITESSE v. 1 is inefficient in large pedigrees with many missing genotypes and can only handle simple pedigrees (no loops and only one set of parents who are founders), the most recent version (v. 2) [11] has implemented a new and faster Elston-Stewart algorithm [12], which computes the conditional probability of a component nuclear family within a pedigree by summing over joint genotypes of the children, instead of the parents (as in v. 1). Furthermore, this summation allows each locus to be computed separately using inheritance vectors analogous to the Lander-Green algorithm [13]. Moreover, a hybrid algorithm is utilized which combines the strength of the Elston-Stewart and Lander-Green algorithm by using a numerical heuristic to decide which technique is best suited for a given nuclear family within a pedigree. Finally, VITESSE v. 2 also has ILINK capabilities and will soon handle pedigrees with loops [11].

Although the LINKAGE/FASTLINK and VITESSE programs remain powerful and popular tools for the analysis of genetic linkage, they are, however, restricted in their multipoint analyses to a limited number of marker loci due to the inherent constraints of the Elston-Stewart algorithm. In short, the Elston-Stewart algorithm scales linearly with 'non-founders' (i.e., individuals whose parents are in the pedigree) and exponentially with loci. Therefore, these programs are best suited to analyze only a small number of marker loci (typically about 1–6, depending on the pedigree size). An alternative to the Elston-Stewart algorithm is the Lander-Green algorithm, which scales exponentially with non-founders and linearly with loci.

The Lander-Green method of extracting complete multipoint linkage data from incomplete marker information (i.e., not all pedigree members are genotyped and/or informative for linkage at each marker) was first implemented in the MAPMAKER program [13]. MAPMAKER implemented an EM [14] search using a hidden Markov model (HMM) to construct maximum likelihood multilocus linkage maps in three-generation Centre d'Etude du Polymorphisme Humain families consisting of four grandparents, two parents and multiple children. In 1995, Kruglyak et al. [15, 16] implemented a speedup in the HMM reconstruction which allowed complete multipoint analysis for homozygosity mapping and linkage analysis in nuclear families (MAPMAKER/HOMOZ) and sibpair analysis (MAPMAKER/SIBS). The MAPMAKER/SIBS program was particularly welcomed by researchers, as it performs model-free multipoint linkage analysis of both qualitative and quantitative traits by examining allele sharing between sibpairs. Allele sharing methods involve studying affected relatives in a pedigree to see how often a particular copy of a chromosomal region (i.e., marker genotypes) is shared. Allele sharing methods do not require a model to be specified for the inheritance of a trait (hence the term 'model-free'), thus making them more suitable for the linkage analysis of

complex diseases – where questions exist regarding disease definitions and the mode of inheritance – may lead to an increase in the chance of making a false rejection of linkage [17]. The MAPMAKER/SIBS program remains a popular (currently being cited over 430 times since 1995) and powerful method for the qualitative and regression-based quantitative analysis of sibpair linkage data.

Utilizing a second speedup of the HMM, the GENEHUNTER package [18] expanded on the method of Kruglyak et al. [16] of extracting complete multipoint data from sibpairs (MAPMAKER/SIBS) by extending it to general pedigrees of modest size (twice the number of nonfounders, 2N – the number of founders, $F \leq 20$). Besides traditional lod score computation, GENEHUNTER included a new (model-free) nonparametric linkage (NPL) statistic, information content mapping and haplotype reconstruction. The NPL analysis is robust to uncertainty about the mode of inheritance, is more powerful than other general pedigree model-free methods (e.g. APM [19] and ERPA [20]) and loses little power relative to traditional lod score analysis (i.e., when there are no errors in the description of the inheritance model) [18].

Since its initial release in 1996, GENEHUNTER v. 1.0 has undergone many improvements. v. 1.1 (motivated by the work of Idury and Elston [21]) and v. 1.2 [22] concerned algorithmic improvements, which substantially increased analysis speed, while v. 1.3 pertained entirely to the X-linked version of GENEHUNTER, correcting problems that had previously been reported by users. Although GENEHUNTER v. 1.0–1.3 concerned significant speedups and bug fixes, it still only calculated a model-based lod or NPL statistic. However, with v. 2.0 came a more unified genetic analysis package incorporating the MAPMAKER/SIBS programs [23], transmission disequilibrium test (TDT) [24] and variance component quantitative trait loci [25] capabilities.

Incorporation of MAPMAKER/SIBS into the GENEHUNTER package (motivated by user requests) allows users to streamline their different analyses by performing them within the same environment, that is, multiple analyses may be obtained within the one run. Similarly, because it has become commonplace for researchers to test for linkage disequilibrium (LD), for example once a linkage has been found and/or to test candidate genes, the inclusion of transmission disequilibrium tests provides users with yet another weapon in the GENEHUNTER arsenal for the analysis of complex traits. In fact, the TDTs implemented in GENEHUNTER are at the cutting edge of LD analysis due to extensions for using missing data, multiple loci *(tdt2, tdt3, tdt4, tdt5)* and estimating significance via a permutation procedure *(perm1, perm2)*. In particular, because a TDT usually involves the analysis of multiple alleles at multiple markers and therefore requires a correction for multiple comparisons, the *perm1* and *perm2* commands, which provide a powerful test of significance (compared to Bonferroni-type corrections), are extremely useful [24].

GENEHUNTER's TDT permutation procedure involves generating a new data set by arbitrarily (with 50% probability) reversing each pair of transmitted and untransmitted alleles, then calculating and storing the TDT results. After this process is repeated many times, as specified by the user (this author recommends using 10,000 replicates), the number of times the permuted data set had a higher TDT result than the real data set (i.e., an empirical p value), and how many of the permuted TDTs had results above a certain threshold (i.e., 0.01, 0.001 and 0.0001) is reported. Therefore, for multiallelic markers with rare alleles – which are unlikely to produce significant results due to their low numbers – the permutation procedure provides an accurate and powerful assessment of significance. Also, by permuting the two-locus haplotype, the *perm2* command adjusts for the non-independence of alleles at two markers in LD with each other, thus providing an accurate estimate of *tdt2* significance. Furthermore, GENEHUNTER allows accurate estimation of the gene effect strength (estimated by transmission ratio) with use of the *dhskip* command. When turned on, *dhskip* eliminates cases in which both parents share the same heterozygous genotype and therefore for whom haplotypes cannot be reconstructed. With *dhskip* turned off, these cases are counted twice, therefore adding half the difference to the result with *dhskip* on produces the corrected test recommended by Markianos et al. [24] and Dudbridge et al. [26].

The recent addition of variance components analysis, which has been shown to be more powerful than regression-based techniques [25, 27–29] for the detection of linkage to quantitative traits, completes GENEHUNTER as a 'one-stop shop' for the statistical genetic analysis of family data. Although other variance component methods using two-point or approximate multipoint data exist, the implementation of variance components analysis within the GENEHUNTER environment offers the added power of an exact (complete) multipoint approach [25]. Although by default GENEHUNTER estimates the variance components separately by sex, it also allows the user (via the *means-by-sex* command) to estimate a single mean when no sex effects are thought to exist, thus increasing analysis speed.

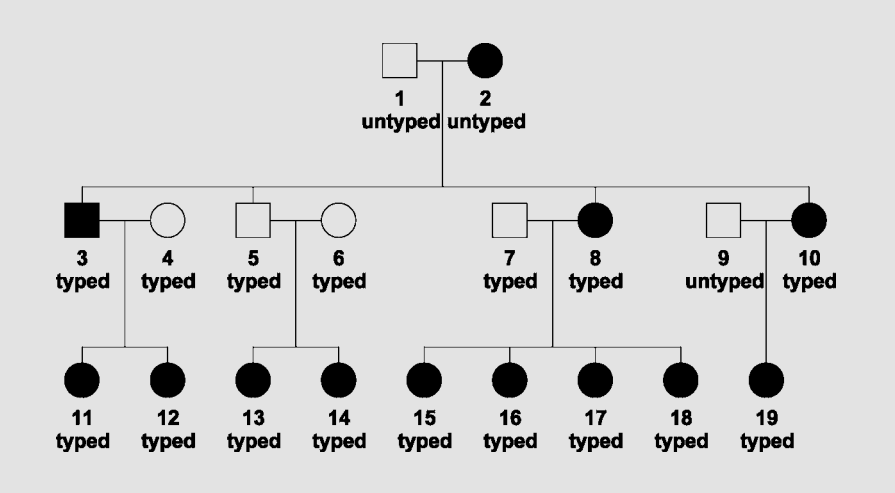

Fig. 1. The 2N – F = 20-bit trimmed migraine pedigree from Nyholt et al. [30].

However, while the speedups of v. 1.1 and v. 1.2 greatly improved the time efficiency of GENEHUNTER, it still required huge amounts of computer processing and memory. For example, using GENEHUNTER v. 2.0, multipoint analysis of 16 markers in a moderately sized migraine pedigree [30], as shown in figure 1, took over 4 days on a dedicated Pentium Pro 200 running LINUX, with 512 Mbytes of RAM and 1 Gbyte of swapspace. Therefore, the most recent improvements to GENEHUNTER v. 2.1 [24], which significantly reduce the computational cost of multipoint analysis by using the observed marker genotypes to restrict the number of potential inheritance configurations, mark a timely and significant advance in the capabilities of GENEHUNTER and its completeness as a package for the analysis of moderately-sized family data. For example, using GENEHUNTER v. 2.1, multipoint analysis of 16 markers in the figure 1 pedigree took only 1 h and 75% less memory than GENEHUNTER v. 2.0. Also, GENEHUNTER v. 2.1 allows users to obtain an estimation of the overall memory requirements and the memory requirements per marker. Therefore, because computational time scales approximately proportional to memory requirements, the user is able to determine whether the computer can support the analysis. To this end, the user may remove (a) marker(s) with significant memory requirements for separate analysis and decide later whether to include them in subsequent multipoint analyses. Moreover, GENEHUNTER also allows the user to turn off *(cs off)* the storage of inheritance probabilities (identity by descent, IBD, matrices) required for the computation of TDT, sibpair statistics and variance components analysis. Hence, although the effect on computational memory is negligible, if such analyses are not required, *cs off* may save considerable disk (storage) memory [24].

Besides the statistical tests aforementioned, GENEHUNTER provides many other useful commands. The *het* argument of the *total stat* command allows the calculation of model-based lod scores under genetic (locus) heterogeneity (i.e., the admixture test [31, 32]). The *count recs* command activates a recombination-counting mechanism, where significant differences between the observed and expected number of recombinations may indicate an error in the map or genotype data. The *haplotype* command will report haplotypes based on the maximum likelihood set of inheritance vectors; by default, the theoretically appealing 'Viterbi' haplotype method is used, however only the 'MaxProb' method *(haplotype method MaxProb)* has benefited from the recent speed improvements; in practice, both methods produce similar results. Therefore, the user should consider using the *haplotype method MaxProb* command on large pedigrees [24]. Lastly, but perhaps of most use, the *dump ibd* command outputs the calculated IBD likelihoods (upon which the above linkage statistics are all based) for each relative pair within each pedigree, which may then be used as input for other programs/analyses; this command expands upon the original *dump ibd* command in MAPMAKER/SIBS, which only outputs IBD values for affected (or quantitatively phenotyped) sibpairs.

Finally, although there have been many improvements and there are clearly many positive aspects of the GENEHUNTER package, it is still somewhat hindered by the size of the pedigree under investigation. Specifically, only

pedigrees of moderate size are able to be analyzed (*max bits* $2N - F \leq 16$ by default). In practice, even on computers with many Gbytes of memory, GENEHUNTER has an upper limit of pedigrees 25–26 bits in size. In fact, due to the representation of inheritance vectors by 32-bit integers, even if one had access to sufficient computational resources, GENEHUNTER is still limited to pedigrees with no more than 32 meioses (i.e., $2N - F \leq 30$ bits). Moreover, as much as the commands *max bits n* (user specification of the pedigree size able to be analyzed), *discard on* (removal of unaffected individuals with no descendants and informative parents) and *skip large off* (pedigrees are trimmed of individuals until it is small enough to be analyzed) allow larger pedigrees to be analyzed using GENEHUNTER, great care must be taken in their use. In fact, as shown by Goedken et al. [33], the use of GENEHUNTER on larger pedigrees (i.e., trimming or splitting – as recommended by Kruglyak et al. [18] – of pedigrees into smaller pedigrees) results in a significant loss of linkage information and/or errors in the estimate of Θ. Moreover, the splitting of pedigrees has the disadvantage of invalidating heterogeneity lod scores [33].

Although there are many other excellent and powerful statistical analysis packages available to the genetic researcher, for example SOLAR [29] (variance components analysis), SIMWALK2 [34] (model-based and model-free linkage and haplotype analysis) and ALLEGRO [35] (many capabilities of GENEHUNTER plus some novel ones, including an independent speedup which, depending on the number of missing genotypes, results in faster analysis speeds than GENEHUNTER v. 2.1) (also see http://linkage.rockefeller.edu/soft/list.html) – in order to provide a concise survey of the GENEHUNTER program – this review has concentrated on only those programs which are most commonly used. Nonetheless, for moderately sized pedigrees GENEHUNTER undoubtedly is the 'gold standard' and a 'one-stop shop' for the statistical genetic analysis of pedigree data. For larger pedigrees, fewer markers must be analyzed at a time, and the LINKAGE, FASTLINK or VITESSE packages remain the programs of choice.

Acknowledgment

Dr. Nyholt is supported by an NHMRC Peter Doherty Postdoctoral Training Fellowship.

References

1 Morton NE: Sequential tests for the detection of linkage. Am J Hum Genet 1955;7:277–318.
2 Ott J: Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am J Hum Genet 1974;26:588–597.
3 Ott J: A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 1976;28:528–529.
4 Lathrop GM, Lalouel JM, Julier C, Ott J: Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 1984;81:3443–3446.
5 Lathrop GM, Lalouel JM: Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 1984;36:460–465.
6 Lathrop GM, Lalouel JM, White RL: Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis. Genet Epidemiol 1986;3:39–52.
7 Cottingham RW Jr, Idury RM, Schaffer AA: Faster sequential genetic linkage computations. Am J Hum Genet 1993;53:252–263.
8 Schaffer AA, Gupta SK, Shriram K, Cottingham RW Jr: Avoiding recomputation in linkage analysis. Hum Hered 1994;44:225–237.
9 Schaffer AA: Faster linkage analysis computations for pedigrees with loops or unused alleles. Hum Hered 1996;46:226–235.
10 O'Connell JR, Weeks DE: The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 1995;11:402–408.
11 O'Connell JR: Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 2001;51:226–240.
12 Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. Hum Hered 1971;21:523–542.
13 Lander ES, Green P: Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 1987;84:2363–2367.
14 Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc (Ser B) 1977;39: 1–22.
15 Kruglyak L, Daly MJ, Lander ES: Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 1995;56:519–527.
16 Kruglyak L, Lander ES: Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 1995;57:439–454.
17 Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 1990;46:222–228.
18 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and non-parametric linkage analysis: A unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.
19 Weeks DE, Lange K: The affected-pedigree-member method of linkage analysis. Am J Hum Genet 1988;42:315–326.
20 Curtis D, Sham PC: Using risk calculation to implement an extended relative pair analysis. Ann Hum Genet 1994;58:151–162.
21 Idury RM, Elston RC: A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered 1997;47:197–202.
22 Kruglyak L, Lander ES: Faster multipoint linkage analysis using Fourier transforms. J Comput Biol 1998;5:1–7.
23 Daly MJ, Kruglyak L, Pratt SC, Houstis N, Reeve-Daly MP, Kirby A, Lander ES: GENEHUNTER 2.0 – a complete linkage analysis system. Am J Hum Genet 1998;63 (suppl): A286.
24 Markianos K, Daly MJ, Kruglyak L: Efficient multipoint linkage analysis. Am J Hum Genet 2001;68:963–977.

25 Pratt SC, Daly MJ, Kruglyak L: Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. Am J Hum Genet 2000;66:1153–1157.
26 Dudbridge F, Koeleman BP, Todd JA, Clayton DG: Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 2000;66:2009–2012.
27 Amos CI, Zhu DK, Boerwinkle E: Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 1996;60:143–160.
28 Amos CI, Krushkal J, Thiel TJ, Young A, Zhu DK, Boerwinkle E, de Andrade M: Comparison of model-free linkage mapping strategies for the study of a complex trait. Genet Epidemiol 1997;14:743–748.
29 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998;62:1198–1211.
30 Nyholt DR, Lea RA, Goadsby PJ, Brimage PJ, Griffiths LR: Familial typical migraine: Linkage to chromosome 19p13 and evidence for genetic heterogeneity. Neurology 1998;50:1428–1432.
31 Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL: The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): Linkage studies, two-locus models, and genetic heterogeneity. Am J Hum Genet 1983;35:1139–1155.
32 Ott J: Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311–320.
33 Goedken R, Ludington E, Crowe R, Fyer AJ, Hodge SE, Knowles JA, Vieland VJ, Weissman MM: Drawbacks of GENEHUNTER for larger pedigrees: Application to panic disorder. Am J Med Genet 2000;96:781–783.
34 Sobel E, Lange K: Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 1996;58:1323–1337.
35 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: Allegro, a new computer program for multipoint linkage analysis. Nat Genet 2000;25:12–13.