Allegro, a new computer program for multipoint linkage analysis

The first generally available computer program for linkage analysis was Liped[1], introduced in 1974. It calculates two-point parametric lod scores for general pedigrees using the Elston-Stewart algorithm[2]. Later programs, including Linkage[3], Fastlink[4] and Vitesse[5], can calculate multipoint lod scores for a few markers, but execution time increases rapidly with the number of markers. Exact multipoint linkage calculations involving many markers for general families was first made practical in 1996 with the introduction of the program Genehunter[6]. The run time of Genehunter grows linearly with the number of markers, but exponentially with pedigree size (p1 in Fig. 1*a* is close to the largest pedigree that can be analysed). In addition to parametric analysis, Genehunter also implemented non-parametric linkage (NPL) scores[6]. These, however, can be conservative when sharing information is incomplete[7,8]; to adjust for this, allele-sharing lod scores based on one-degree-of-freedom models were introduced[7] and implemented in a modified version of Genehunter, Genehunter-Plus.

We have developed several improvements to the computational algorithms of Genehunter. The new algorithms have been incorporated in a computer program for multipoint linkage analysis, Allegro, which is available free for non-commercial use (e-mail: allegro@decode.is). Included is a program manual, a technical report and the source code. Like Genehunter, Allegro uses a hidden Markov model[9–11], and time and memory costs still grow exponentially with pedigree size, but it is considerably faster than Genehunter (typically 20–100 times, Table 1). Apart from allowing for larger pedigrees (typically 20–30% larger), the speed improvement is relevant for simulation studies. Instead of providing the technical details of the computational algorithms, here we describe Allegro from the user's perspective.

Allegro has much of the functionality of Genehunter and Genehunter-Plus. Specifically, Allegro calculates multipoint parametric lod scores, NPL scores and allele-sharing lod scores based on the scoring functions S_{pairs} and S_{all} (ref. 12), reconstruction of haplotypes, estimated recombination count between markers (observed map), and entropy information. The X chromosome is supported in all calculations. Although planned for the future, Allegro currently does not support quantitative traits.

For non-parametric analysis, Genehunter gives a *P* value computed by comparing the observed NPL score with its complete data distribution. This *P* value is usually conservative when the sharing information is incomplete. Allegro gives this *P* value and another *P* value that is calculated by comparing the observed allele-sharing lod score with its complete data distribution. This latter *P* value is probably more accurate in most situations, but is not guaranteed to be conservative. When the observed data provide complete sharing information, the two *P* values coincide. Genehunter weights the standardized forms of the family scores equally when tabulating the overall score, potentially reducing power when the families vary in size by giving too much weight to small families. When using S_{pairs}, an alternative is to weight all pairs equally, but this might put too much weight on large families. For example, a proposed scheme[13] downweights affected pairs in sibships which have more than two affected sibs, which falls in-between weighting the families equally and the pairs equally. Allegro supports this and other strategies[14] by allowing the user to specify the weighting scheme. This feature can also be used to perform conditional analyses in which the weight of a family depends on some linkage score or specific genotypes of the family at another locus[15].

In addition to S_{pairs} and S_{all}, and the entropy measure, Allegro supports other scoring functions[16] and information measures[17]. Other advantages are improved input and output, and the ability to perform multiple analyses together at little extra cost (with different parametric models, scoring functions and family-weighting schemes). Also, at a cost of 10–30% in run time, Allegro will, if necessary, use recalculation or disk-swapping to cut down memory requirements by a factor of 20–60 compared with Genehunter. Analysing pedigree p3 (Fig. 1) for 50 markers requires about 900 megabytes of memory using Genehunter, but less than 15 megabytes using Allegro.

Allegro can simulate multi-locus marker data either under no linkage or under linkage. Simulations under no linkage can be used to study the calibration of

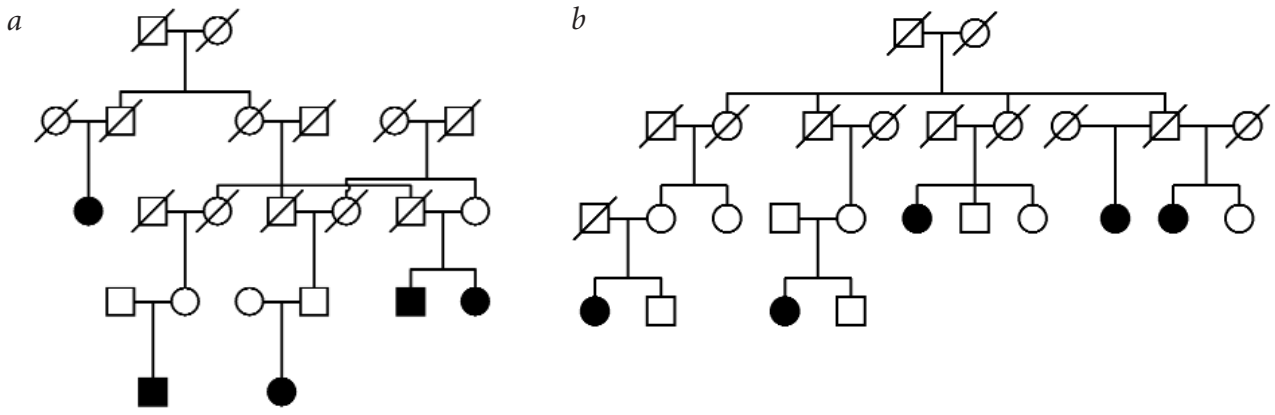

Fig. 1 Pedigrees used in timing experiments. Filled symbols represent affected individuals and genotypes are unavailable for individua ls represented by symbols with a slash. ***a***, Pedigree p1, containing a simple marriage loop. ***b***, Pedigree p4, which is approximately the largest pedigree that can be analysed with Allegro on a computer with one gigabyte of memory. Pedigree p3 is obtained from pedigree p4 by removing the seventh, eighth and eleventh individuals in the third generation, and pedigree p2 is obtained by further removing the remaining unaffected sister in the third generation and the unaffected brothers in the fourth generation.

Table 1 • Run times for Genehunter (version 1.3) and Allegro

	Without haplotyping			With haplotyping		
Pedigree	Genehunter	Allegro	Factor	Genehunter	Allegro	Factor
p1	67 min	0.68 min	99	159 min	1.53 min	104
p2	46 s	1.11 s	41	100 s	2.23 s	45
p3	44 min	1.17 min	38	137 min	2.91 min	47
p4	NA	241 min	NA	NA	487 min	NA

Results from timing experiments on a Sun Ultra Enterprise 450 computer with 2 gigabytes of memory. Each run consists of one parametric and one non-parametric analysis for 50 markers. For these analyses version 1.3 is the fastest version of Genehunter available. The current version (2.0) is about twice as slow, because it automatically computes posterior IBD sharing probabilities for all pairs of relatives, a feature that is not incorporated in version 1.3. Allegro provides this feature as an option, and for these pedigrees the run time increases by about 50%. This means that the factors of improvement of Allegro with this option active compared with Genehunter 2.0 are greater than the numbers in the table. Genehunter cannot handle pedigrees larger than pedigrees p1 and p3 on this computer. If one or both of the founders in p2–p4 were genotyped, Allegro cannot take advantage of symmetry that exists between the founders; the run-time is doubled.

different methods of calculating *P* values. One can determine genome-wide adjusted *P* values for specific families and marker density. Under linkage to a susceptibility gene, Allegro will simulate marker data conditional on the observed disease phenotypes and a given inheritance model[18]. These simulations can be used to assess the power of a set of families, or the effects of marker density and missing data. They can also be used to compare parametric and non-parametric methods, various scoring functions and weighting schemes, and SNPs versus microsatellites.

Acknowledgements

We thank the statistics group at Decode Genetics, D. Nicolae and H. Cordell for extensive testing of Allegro and suggestions.

Daniel F. Gudbjartsson[1,2], Kristjan Jonasson[1], Michael L. Frigge[1] & Augustine Kong[1,3]

[1]Decode Genetics, Lynghals 1, 110 Reykjavik, Iceland. [2]Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, USA. [3]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. Correspondence should be addressed to D.F.G. (e-mail: dfg@decode.is).

1. Ott, J. *Am. J. Hum. Genet.* **26**, 588–597 (1974).
2. Elston, R.C. & Stewart, J. *Hum. Hered.* **21**, 523–542 (1971).
3. Lathrop, G.M., Lalouel, J.M., Julier, C. & Ott, J. *Proc. Natl Acad. Sci. USA* **81**, 3443–3446 (1984).
4. Cottingham, R.W., Idury, R.M. & Schäffer, A.A. *Am. J. Hum. Genet.* **53**, 252–263 (1993).
5. O'Connell, J.R. & Weeks, D.E. *Nature Genet.* **11**, 402–408 (1995).
6. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
7. Kong, A. & Cox, N.J. *Am. J. Hum. Genet.* **61**, 1179–1188 (1997).
8. Badner, J.A., Gershon, E.S. & Goldin, L.R. *Am. J. Hum. Genet.* **63**, 880–888 (1998).
9. Lander, E.S. & Green, P. *Proc. Natl Acad. Sci. USA* **84**, 2263–2367 (1987).
10. Kruglyak, L. & Lander, E.S. *J. Comput. Biol.* **5**, 1–7 (1998).
11. Idury, R.M. & Elston, R.C. *Hum. Hered.* **47**, 197–202 (1997).
12. Whittemore, A.S. & Halpern, J. *Biometrics* **50**, 118–127 (1994).
13. Hodge, S.E. *Genet. Epidemiol.* **1**, 109–122 (1984).
14. Weeks, D.E. & Lange, K. *Am. J. Hum. Genet.* **42**, 315–326 (1988).
15. Cox, N.J. *et al*. *Nature Genet.* **21**, 213–215 (1999).
16. McPeek, M.S. *Genet. Epidemiol.* **16**, 225–249 (1999).
17. Nicolae D.L. *Allele Sharing Models In Gene Mapping: A Likelihood Approach*. Thesis, Univ. Chicago (1999).
18. Ploughman, L.M. & Boehnke, M. *Am. J. Hum. Genet.* **44**, 543–551 (1989).