AUTOGSCAN: Powerful Tools for Automated Genome-Wide Linkage and Linkage Disequilibrium Analysis

Tero Hiekkalinna,[1] Joseph D. Terwilliger,[4,5,6] Sampo Sammalisto,[1,3] Leena Peltonen,[1,2,3] and Markus Perola[1,3]

1 Department of Molecular Medicine, National Public Health Institute, Helsinki, Finland

2 Department of Human Genetics, David Geffen School of Medicine at the University of California, Los Angeles, California, United States of America

3 Department of Medical Genetics, University of Helsinki, Helsinki, Finland

4 Finnish Genome Center, University of Helsinki, Helsinki, Finland

5 Columbia Genome Center, Department of Psychiatry and Department of Genetics and Development, Columbia University, New York, United States of America

6 Division of Molecular Genetics, New York State Psychiatric Institute, New York, United States of America

Genome-wide linkage analysis using multiple traits and statistical software packages is a tedious process which requires a significant amount of manual file manipulation. Different linkage analysis programs require different input file formats, making the task of analyzing data with multiple methods even more time-consuming. We have developed a software tool, AUTOGSCAN, that automates file formatting, the running of statistical analyses, and the summarizing of resulting statistics for whole genome scans with a push of a button, using several independent, and often idiosyncratic, statistical software packages such as MERLIN, SOLAR and GENEHUNTER. We also describe a program, ANALYZE, designed to run qualitative linkage analysis with several different statistical strategies and programs to efficiently screen for linkage and linkage disequilibrium for a given discrete trait. The ANALYZE program can also be used by AUTOGSCAN in a genome-wide sense.

The use of linkage and association analysis programs in genome-wide analysis typically requires a large amount of file manipulation, especially when a variety of different idiosyncratic software packages are used. This basically trivial repetitive process can be cumbersome, time-consuming and error-prone because the various software packages have little uniformity in formatting standards. Since a genome scan normally requires the user to repeat the basic steps for file manipulation and collation of results multiple times, automation has obvious advantages, and since files needed for the statistical analyses can be both large and numerous, manual manipulations are rarely advisable. We have developed software to automate the tedious process of creating, modifying and interpreting input and output files for various software packages, which allows efficient analyses of several different phenotypic traits under a variety of statistical models with the push of a button. In practice, these programs enable the analyses of a genome-wide scan for all markers on all chromosomes with a variety of linkage and family-based association analysis programs in a single step. The AUTOGSCAN software package applies a variety of statistical tools for two-point and multipoint linkage and the association analysis of a discrete or continuous trait, using the following analysis software: FASTLINK 4.1P (Cottingham et al., 1993; Schaffer et al., 1994) version of the LINKAGE package (Lathrop & Lalouel, 1984; Lathrop et al., 1984; Lathrop et al., 1986), SIBPAIR (Göring & Terwilliger, 2000b; Kuokkanen et al., 1996), HRRLAMB (Terwilliger, 1995; Terwilliger & Ott, 1992), HOMOG (Ott, 1985), GENEHUNTER (Kruglyak et al., 1996; Kruglyak & Lander, 1998), MERLIN (Abecasis et al., 2002) and SOLAR (Almasy & Blangero, 1998). Much emphasis has been placed on standardization and simplification of the requisite input files and program commands to ease the analysis of a dataset with multiple phenotypes and multiple software packages including simplifying and summarizing the results. The only input files required for these automation programs are LINKAGE format pedigree files (one file per chromosome, with all markers in chromosome order), marker map files (defining the distances between consecutive markers on a given chromosome), phenotype description file(s) and, when needed, additional files defining specific details on how each phenotype is to be analyzed.

Received 26 November, 2004; accepted 3 December, 2004.

Address for correspondence: Leena Peltonen, MD, PhD, Academy Professor, National Public Health Institute, Department of Molecular Medicine, PO Box 104, 00251 Helsinki, Finland.
E-mail: leena.peltonen@ktl.fi

Methods

Input File Formats

AUTOGSCAN requires standard input file formats: LINKAGE format 'pedigree files' (Terwilliger & Ott, 1994) for all markers on each chromosome (defining the pedigree structure, affection status phenotype, if any, and marker locus genotypes), MEGA2 (Mukhopadhyay et al., 1999) format 'map files' for each chromosome (chromosome number, marker positions along the chromosome in Haldane cM, and marker names) and trait locus 'phenotype files' where needed, which include quantitative trait and optional covariate values for each phenotyped individual. Input file requirements are listed in Table 1 and program input–output flow is illustrated in Figure 1. Detailed information for each input file format as well as working example files are provided on the web page <http://www.helsinki.fi/~tsjuntun/autogscan/index.html>.

PEDCONVERT

PEDCONVERT provides the core for the automation software; it handles pedigree and map file conversion and input file creation for each analysis package. PEDCONVERT uses red-black tree data structures (Rudolf, 1972), which are a variant of self-balancing binary search trees, for efficiently merging pedigree and phenotype files into a single file with marker locus genotypes and trait locus phenotypes and additional trait locus covariates. Both pedigrees and individuals within a pedigree are stored in red-black tree data structures. This enables insertion of phenotypes for each individual in logarithmic time (O[log n], Knuth, 1976, where n is the number of individuals in the tree), which significantly decreases PEDCONVERT run time for large datasets. The run time comparison between red-black trees and linked-list data structures (O[n]) when joining whole genome-wide pedigree data (23 chromosomes and 400 marker loci) and phenotype files with different sample sizes are shown in Table 2. The files created by PEDCONVERT for different analysis programs described below remain in directories after the analyses and can be used by the investigator for further analysis.

Discrete Trait Two-Point Analysis

Two-point linkage and association analysis for discrete traits is performed using the following set of programs incorporated into a software control package: ANALYZE — Parametric linkage analysis using MLINK from FASTLINK 4.1P (Cottingham et al., 1993; Schaffer et al., 1994), homogeneity testing using HOMOG (Ott, 1985), pseudomarker affected sib-pair analysis using SIBPAIR (Kuokkanen et al., 1996) and family-based association testing by using Haplotype Based Haplotype Relative Risk (HHRR; Terwilliger & Ott, 1992). Required files for these analyses include pedigree files (linkage format) and model files, which describe the mode of inheritance for parametric linkage analysis of discrete traits, in the standard LINKAGE 'Affection status locus' format (see Terwilliger & Ott, 1994 for specifics). Optional loop files are needed for breaking loops, if any are present in the pedigrees, in 'MAKEPED' format (Terwilliger & Ott, 1994). The following steps are done automatically after launching the program: (1) pedigree files are processed using the MAKEPED program (Terwilliger & Ott, 1994); (2) a LINKAGE formatted parameter file is created; and (3) allele frequencies are estimated from the data with a simple gene-counting algorithm with DOWNCODE, included in the package (Göring & Terwilliger, 2000a) and estimated allele frequencies are inserted in the parameter file. DOWNCODE further eliminates missing alleles, and renumbers the alleles consecutively starting with 1 prior to analysis. The user has the option to check file integrity before starting more time-consuming statistical analyses. However, in all cases, before any analysis is performed, the PEDCHECK program (O'Connell & Weeks, 1998) is automatically used to identify any Mendelian inconsistencies in the pedigree data. If such inconsistencies are found, that particular chromosome is censored from the analysis until all inconsistencies are eliminated. Finally, two-point linkage and association analyses are performed for each marker on each chromosome consecutively.

Table 1

Input File Requirements for Each of the Auto-Programs and Default Analysis Methods Used

Program	Default methods	Pedigree files	Marker map files	Phenotype file	Model file	Control file
AUTOQUALITATIVE	Parametric and nonparametric two-point linkage analysis for binary traits	x		x (optional)	x	x (optional)
AUTOGENEHUNTER	Parametric and nonparametric two-point and multipoint linkage analysis for binary traits	x	x	x (optional)	x	x (optional)
AUTOMERLIN	VC and NPL linkage analyses, MIBD estimation, genotyping error detection	x	x	x		
AUTOSOLAR	VC linkage analysis	x	x	x		x

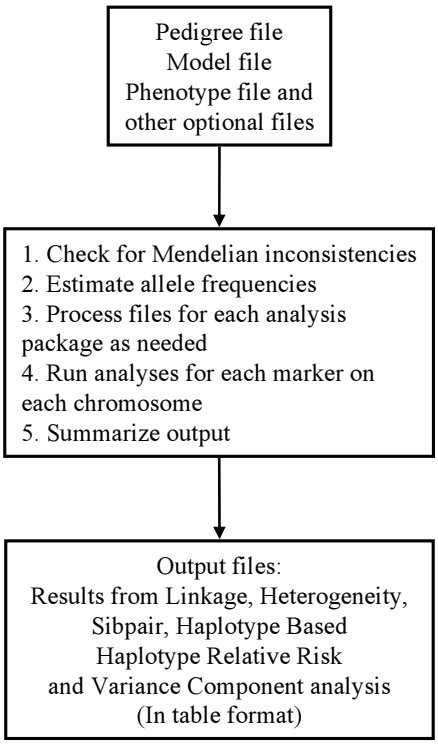

Figure 1

Input and output file flow in the AUTOGSCAN program and analysis steps performed for each of the chromosomes.

Output files for each chromosome consist of both detailed analysis files (produced by the original software packages themselves) and summary files, with the statistics (not the associated parameters) presented in a table for all markers. To ease the processing of genome-wide results, all summary files from all chromosomes are saved to one text file as well. Reanalysis with different modes of inheritance requires changing the model file only. The program assumes that the first locus in the file is the 'affection status' trait locus, but in a separate phenotype file the user may indicate multiple additional discrete traits for testing. In this case, PEDCONVERT would be used to concatenate pedigree and phenotype files repeatedly as needed. The program accepts multiple additional command line parameters, for example, which analyses to perform, with detailed enumeration of all potential command line parameters listed on the auto-programs web page.

A stand-alone version of this program, ANALYZE, which controls the statistical analyses described above (including also multipoint association analysis [Terwilliger, 1995]), is also included. This program can be used for analyzing qualitative trait data using one pedigree and one locus file without using the larger automation program. The program requires LINKAGE format pedigree and a locus file. Output files contain detailed information and a summary of performed analyses.

Multipoint Linkage Analysis

Model-based and model-free linkage analysis with a discrete trait can also be performed using the GENEHUNTER or MERLIN (nonparametric option) programs. MERLIN requirements are described in the 'Quantitative trait analysis' section below. The information required for this analysis is the same as above, except that no loop file is needed as GENEHUNTER processes loops differently to LINKAGE, and additional marker map information is needed for each chromosome. Again, the following steps are automatically performed after the initiation of the program: first, Mendelian inconsistencies are checked in the pedigree file with PEDCHECK; second, a LINKAGE format locus file is created as above; third, allele frequencies are estimated from the data using DOWNCODE. The program then creates a batch file, which enables the running of GENEHUNTER in batch mode, that is, noninteractive mode. The batch file includes all relevant GENEHUNTER commands, loading locus file information, setting up marker map and so on, all of which are required for analysis. Finally, GENEHUNTER-based analysis is performed. This procedure is repeated for all chromosomes. Output files for each chromosome are two-point and multipoint text files which are saved to one two-point and one multipoint text file, postscript graph files for parametric and nonparametric multipoint curves and information content curves. If the user should opt to reanalyze the data under different assumptions in relation to the mode of inheritance, the user would need only to modify the model and phenotype files. As above, the program accepts several command line parameters for controlling automated analysis and all parameters are listed on the AUTOGSCAN web page.

Quantitative Trait Analysis

Two-point and multipoint variance components (VC) quantitative trait locus (QTL) linkage analysis is performed using the MERLIN or SOLAR programs. However, in order to provide the user with comprehensive access to the wide range of options implemented in MERLIN, any of the MERLIN analysis options (e.g., nonparametric analysis [NPL], haplotype analysis) can be performed by using 'merlin-override' files; detailed information regarding this can be found on the previously mentioned web site. The information required for all analyses consists of LINKAGE format pedigree files, MEGA2 format map files and a phenotype file. First, the pedigree file and phenotype file are combined with PEDCONVERT, which also creates MERLIN format locus and map files, and analyses are performed subsequently for each chromosome. The program provides an option to check for Mendelian inconsistencies and review trait components by using the PEDSTATS program which is part of the MERLIN package. Output files for each chromosome in VC analysis represent the standard MERLIN two-point and multipoint text files as well as optional MERLIN

Table 2:
Average PEDCONVERT Run Time Comparisons

Number of pedigrees	Number of individuals in a pedigree	Total number of concatenations	Time: red-black tree	Time: linked-list
50	4	4600	4.7 s	5.0 s
100	4	9200	9.5 s	11.2 s
200	4	18400	19.5 s	29.7 s
300	4	27600	28.8 s	1 m 3 s
400	4	36800	38.0 s	1 m 47 s
500	4	46000	46.9 s	2 m 40 s
1000	4	92000	1 m 31 s	9 m 41 s

Note: The average PEDCONVERT run time (reading pedigree and phenotype file and forming the new files) comparisons when pedigrees were stored to red-black tree or linked-list data structures on a 2 GHz AMD Opteron with 12GB of RAM running 64-bit Fedora Core 2 Linux operating system. Note that the time depends also on hard drive read and write speed (disc I/O).

generated PDF graph files for two-point and multipoint curves. In addition, all chromosomal two-point and multipoint text files are saved in the form of one two-point and one multipoint text file for easy viewing of results. The number of output files depends on the selected analysis option. Detailed information for all the command line parameters is accessible on the AUTOGSCAN web page.

This package may also be used to perform two-point and multipoint variance components QTL linkage analysis with the SOLAR package. The input files required are the same as for AUTOMERLIN, although one additional SOLAR control file is needed. This additional file is used to specify which trait or traits are to be analyzed, and which will be used as covariates in the VC analysis along with other SOLAR specific commands that the user may wish to use. First, Mendelian checking for inconsistencies is performed for each chromosome by using PEDCHECK. Second, pedigree, map and phenotype files are converted to SOLAR format by using PEDCONVERT. Finally the analysis is performed. The same steps are performed for each chromosomes. It is strongly recommended that two-point and multipoint IBD (Identity By Descent) matrices be calculated prior to linkage analysis using a control file, since these need to be calculated for all trait analyses only once, except when the genotype data changes. Output files are two-point and multipoint text files for each chromosome. SOLAR also creates several other output files, which are stored in the analysis folder. All additional command line parameters information, as well as how to calculate IBD matrices, are located in the AUTOGSCAN web page.

Discussion

Genome-wide linkage analysis of multiple traits requires tedious input file manipulation when performed by hand without specialized automation programs. This tedious file processing introduces errors to the data files and is extremely time-consuming. Automation programs have been developed to handle these procedures and made available to other scientists so that valuable research time is not wasted on the trivial repetitive tasks of file manipulation and the learning of various analysis programs' idiosyncrasies. Since automation reduces data-input errors, it is believed that these programs will be extensively used in gene-mapping efforts, thus providing accurate results from the simple 'first-pass' analyses with less manual labor. In practice, more sophisticated methods are typically warranted in the final analysis for which task professional statistical geneticists are invaluable; however, there is no need for professionals to spend time on the repetitive summarizing of genome scan information for basic and simple routine analyses. Programs like AUTOGSCAN are therefore of utmost value in an age where the demand exceeds the supply of trained professionals. It is clear that automation will become a necessity in the very near future when high-throughput SNP genotyping platforms begin to yield datasets of tens if not hundreds of thousands of marker genotypes, although the utility and power of such approaches for linkage analysis remains in doubt (Hiekkalinna et al., 2004).

There is a danger, however, that although these programs reduce the time needed to start the analyses for a total genome scan to a few seconds after a set of files has been produced, they will certainly not reduce the time needed to review the results and make scientifically valid conclusions from them. It is of utmost importance that all genome-wide tool users know and understand the original computer programs as well as the assumptions made by the analysis methods implemented. Users would then be aware of the potential benefits, limitations and pitfalls for each program, as well as the justifications for selecting one program over another. In addition, it has already been noticed in beta-testing that ease of analyses does tempt users to perform multiple testing unnecessarily, and the importance of a basic scientific rule should indeed be applied and emphasized: the careful a priori planning of any analyses is essential to the interpretation.

However, it is thought that the time saved by using such programs provides more time for such efforts as well as other quality issues. To conclude, it should be emphasized that the programs described here are intended for the efficient initial screening of available genome-wide data using fixed options. It should be noted that the default options of these programs may not be optimal for all data sets, and that it is therefore necessary to consider a priori overriding default options as well as utilizing the full variety of analysis options implemented in the actual analysis programs for interesting loci detected in initial screens.

All binaries and source codes (written in Unix Shell Script [BASH], C and C++) presented here are freely available for academic use and can be downloaded from the web page <http://www.helsinki.fi/~tsjuntun/autogscan/index.html>. This web page has complete installation instructions and links for the actual analysis programs, which must be obtained from the primary sources. It is only the automating and summarizing software that is distributed by the authors, in addition to the analysis software, SIBPAIR and HRRLAMB.

Acknowledgments

We wish to thank the developers of the programs PEDCHECK, MEGA2, MERLIN, SOLAR and GENEHUNTER for providing these valuable assets to the scientific community. The authors would like to emphasize that publications describing the original programs should also be cited when publishing results where the authors' tools are used. Juha Knuuttila's valuable comments and feedback for the auto-programs and manuscript are appreciated. Funding from the National Institutes of Health (Grants NS43559 and HL70150-01A1 for LP and Grant MH63749 to JDT) is gratefully acknowledged, along with the support from the Center of Excellence in Disease Genetics of the Academy of Finland and an Emil Aaltonen Foundation young scientist grant for T.H. This work has been done with support from the GenomEUtwin project which is supported by the European Commission under the program 'Quality of Life and Management of the Living Resources' of 5th Framework Programme (no. QLG2-CT-2002-01254), Biocentrum Helsinki and Helsinki Biomedical Graduate School (S.S.).

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin: Rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, *30*, 97–101.

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, *62*, 1198–1211.

Cottingham, R. W., Jr., Idury, R. M., & Schaffer, A. A. (1993). Faster sequential genetic linkage computations. *American Journal of Human Genetics*, *53*, 252–263.

Göring, H. H., & Terwilliger, J. D. (2000a). Linkage analysis in the presence of errors III: Marker loci and their map as nuisance parameters. *American Journal of Human Genetics*, *66*, 1298–1309.

Göring, H. H., & Terwilliger, J. D. (2000b). Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *American Journal of Human Genetics*, *66*, 1310–1327.

Hiekkalinna, T. S., Perola, M., & Terwilliger, J. D. (2004). Linkage analysis genome scans with tens of thousands of SNPs gives systematically higher power than traditional microsatellite-based approaches under the NULL hypothesis [Abstract]. *Twin Research*, *7*, 683.

Knuth, D. E. (1976). Big omicron and big omega and big theta. *ACM SIGACT News*, *8*, 18–23.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics*, *58*, 1347–1363.

Kruglyak, L., & Lander, E. S. (1998). Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, *5*, 1–7.

Kuokkanen, S., Sundvall, M., Terwilliger, J. D., Tienari, P. J., Wikstrom, J., Holmdahl, R., Pettersson, U., & Peltonen, L. (1996). A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2. *Nature Genetics*, *13*, 477–480.

Lathrop, G. M., & Lalouel, J. M. (1984). Easy calculations of lod scores and genetic risks on small computers. *American Journal of Human Genetics*, *36*, 460–465.

Lathrop, G. M., Lalouel, J. M., Julier, C., & Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences, USA*, *81*, 3443–3446.

Lathrop, G. M., Lalouel, J. M., & White, R. L. (1986). Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis. *Genetic Epidemiology*, *3*, 39–52.

Mukhopadhyay, N., Almasy, L., Schroeder, M., Mulvihill, W. P., & Weeks, D. E. (1999). Mega2: A data-handling program for facilitating genetic linkage and association analyses [Abstract]. *American Journal of Human Genetics*, *65*, A436.

O'Connell, J. R., & Weeks, D. E. (1998). PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, *63*, 259–266.

Ott, J. (1985). *Analysis of human genetic linkage*. Baltimore, MD: Johns Hopkins University Press.

Rudolf, B. (1972). Symmetric Binary B-Trees: Data Structure and Maintenance Algorithms. *Acta Informatica*, *1*, 290–306.

Schaffer, A. A., Gupta, S. K., Shriram, K., & Cottingham, R. W., Jr. (1994). Avoiding recomputation in linkage analysis. *Human Heredity*, *44*, 225–237.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, *56*, 777–787.

Terwilliger, J. D., & Ott, J. (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity*, *42*, 337–346.

Terwilliger, J. D., & Ott, J. (1994). *Handbook of human genetic linkage*. Baltimore, MD: Johns Hopkins University Press.