SOURCE CODE FOR BIOLOGY AND MEDICINE

SOFTWARE REVIEW Open Access

Mega2: validated data-reformatting for linkage and association analyses

Robert V Baron[1], Charles Kollar[1], Nandita Mukhopadhyay[2] and Daniel E Weeks[1,3*]

Abstract

Background: In a typical study of the genetics of a complex human disease, many different analysis programs are used, to test for linkage and association. This requires extensive and careful data reformatting, as many of these analysis programs use differing input formats. Writing scripts to facilitate this can be tedious, time-consuming, and error-prone. To address these issues, the open source Mega2 data reformatting program provides validated and tested data conversions from several commonly-used input formats to many output formats.

Results: Mega2, the Manipulation Environment for Genetic Analysis, facilitates the creation of analysis-ready datasets from data gathered as part of a genetic study. It transparently allows users to process genetic data for family-based or case/control studies accurately and efficiently. In addition to data validation checks, Mega2 provides analysis setup capabilities for a broad choice of commonly-used genetic analysis programs. First released in 2000, Mega2 has recently been significantly improved in a number of ways. We have rewritten it in C++ and have reduced its memory requirements. Mega2 now can read input files in LINKAGE, PLINK, and VCF/BCF formats, as well as its own specialized annotated format. It supports conversion to many commonly-used formats including SOLAR, PLINK, Merlin, Mendel, SimWalk2, Cranefoot, IQLS, FBAT, MORGAN, BEAGLE, Eigenstrat, Structure, and PLINK/SEQ. When controlled by a batch file, Mega2 can be used non-interactively in data reformatting pipelines. Support for genetic data from several other species besides humans has been added.

Conclusions: By providing tested and validated data reformatting, Mega2 facilitates more accurate and extensive analyses of genetic data, avoiding the need to write, debug, and maintain one's own custom data reformatting scripts. Mega2 is freely available at https://watson.hgen.pitt.edu/register/.

Keywords: Software, Linkage, Association, Human Genetics, Data management

Background

The gene-discovery process is very well advanced at the data-generation end with sophisticated database management systems, laboratory information management systems, and bioinformatics tools. There has also been enormous progress in terms of analytical software. However, very little has been done to facilitate the efficient transfer of data from the generation stage to the analysis stage; analysis programs have diverse and stringent requirements (not always clearly documented) on how the input data should be formatted, which is often very different from how the generated data are formatted. Researchers face the need to collect and collate genetic data from diverse sources, and this need has increased significantly as rapidly improving technology generates orders of magnitude more data. As new analysis programs come into being, data setup and organization continues to be an error-prone and very time-consuming task if performed manually, but ideal for well-tested computer automation.

In the course of a single study of the genetics of a complex disease, the optimal analysis might require use of several different programs. For example, one might want to use pedstats [1] to check for data validity, PREST [2,3] to check for relationship errors, SOLAR [4,5] to test for linkage, and Mendel [6-8] to test for association in the presence of linkage. Each provides the best possible analysis but also has its own strict input format requirements, so

* Correspondence: weeks@pitt.edu
[1]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA
[3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA
Full list of author information is available at the end of the article

there is great value in being able to quickly and easily convert one's data format as required.

To meet these needs, we developed Mega2, the Manipulation Environment for Genetic Analysis [9,10], which automates common data reformatting tasks, thereby accelerating analyses, saving time, and reducing errors. We describe here recent major updates to Mega2, which include improvements in memory efficiency, improved support for commonly used input formats such as PLINK and VCF, and addition of several more target output formats.

Implementation

Mega2 was originally released in January 2000, and has undergone continuous revisions since. Mega2 was originally written in C, but now has been written in C++, allowing us to now use modern object-oriented programming techniques. Mega2 was designed to be used in a Unix environment, and so for extended functionality, such as plotting results with R or running generated scripts, uses a few other programs commonly available in the Unix environment, such as Perl, Awk, Python, tcsh and bash-shells, and R. Perl is used for producing formatted output such as tables and HTML reports, and R is used to create graphical output using our R "nplplot" package. The currently released version (4.7.1) of Mega2 is available in Additional file 1; for updated versions, please visit the project home page as listed in the "Availability and requirements" section.

The current Mega2 implementation transforms a matched set of pedigree, phenotype, genotype, and map (genetic and physical) information into a matched set of output files that are ready for analysis by one of many commonly-used genetic analysis tools. Accordingly, Mega2 is organized into input, error-checking, reordering, and output components, constituting a single-layered architecture, and communicating directly with each other as necessary. Mega2 is a command-line program that is typically run without arguments in an interactive mode, where a sequence of menus (Figure 1, blue blocks) are presented to the user to specify 1) input files and filters, 2) the target analysis program, 3) program-specific options, 4) plot customization options, 5) the subset of loci to be included in the output data, and 6) the trait loci and covariates. The input data can be in LINKAGE format [11-13], Mega2 annotated format, PLINK (ped or binary) format [14], or VCF/BCF format [15]. When reading from Mega2 annotated format, input files include a) a pedigree file containing sample-related pedigree, phenotype, and genotype information, b) a locus names file, and c) a map file containing chromosomal positions for marker loci. Additionally, the user may specify d) an omit file for setting selected genotypes to unknown, e) an allele-frequency file, and f) a disease-model or penetrance file. Mega2 reads (Figure 1, orange circle) and validates the input data, and creates output files to be run by the target program, possibly partitioned by chromosome and/or traits. Output usually includes (Figure 1, green blocks) scripts to run the analysis on the re-organized, re-formatted data, and may create plots of the analysis results, as well as BED custom tracks suitable for plotting in the UCSC Genome Browser, using our 'nplplot' R library. Detailed diagnostic measures are created in certain cases. Full logs of run-related information including error messages are generated. Mega2 can also be run in a non-interactive mode using a batch file (Figure 1, top center) containing (key, value) pairs corresponding to the choices made via the interactive menus.

Results

Mega2 was originally written without much attention to memory efficiency, as at that time a genome-wide scan consisted of only several hundred markers. Thus, Mega2's memory usage was initially on the order of people x allele x 8 bytes, as each person/allele combination was assigned a pointer to the allele label. For two-allele marker data, we have markedly reduced memory requirements by replacing each pair of 8 byte pointers with a 2 bit index specifying which alleles the individual has. We also allow the user to switch out of the 2 bit mode if they want to work with more highly polymorphic markers. Further memory efficiencies have been gained by not storing (unknown) genotypes for completely untyped individuals, but who are still needed to specify the pedigree structure. As a result of these improvements, Mega2 now can handle genome-wide scale data – for example, 895K two-allele markers on 3.1K people requires only 1.12 Gb of memory for Mega2 processing.

Mega2 can now read data in from a wider variety of input formats. Many researchers now have their data in PLINK-format [14], so we have extended Mega2 to support reading PLINK input files. Mega2 now directly processes PLINK 'ped' and binary input formats. Mega2 also supports PLINK phenotype files, as well as Mega2-format map files that specify a sex-specific genetic map. Furthermore, we recently added support for reading Variant Call Format (VCF) files and their binary compressed equivalent, BCF; most sequencing-based data are now in VCF/BCF format [15].

As can be seen in Table 1, Mega2 currently reformats data for a wide variety of target programs that perform linkage and association analysis, a popular pedigree-drawing program Cranefoot [16], as well as others that perform quality-control analyses such as computing genotyping success rates, testing for departures from Hardy-Weinberg equilibrium, etc. Mega2 is now able to generate PLINK binary output files. As data sets get larger, Mega2's support of the PLINK binary format, both

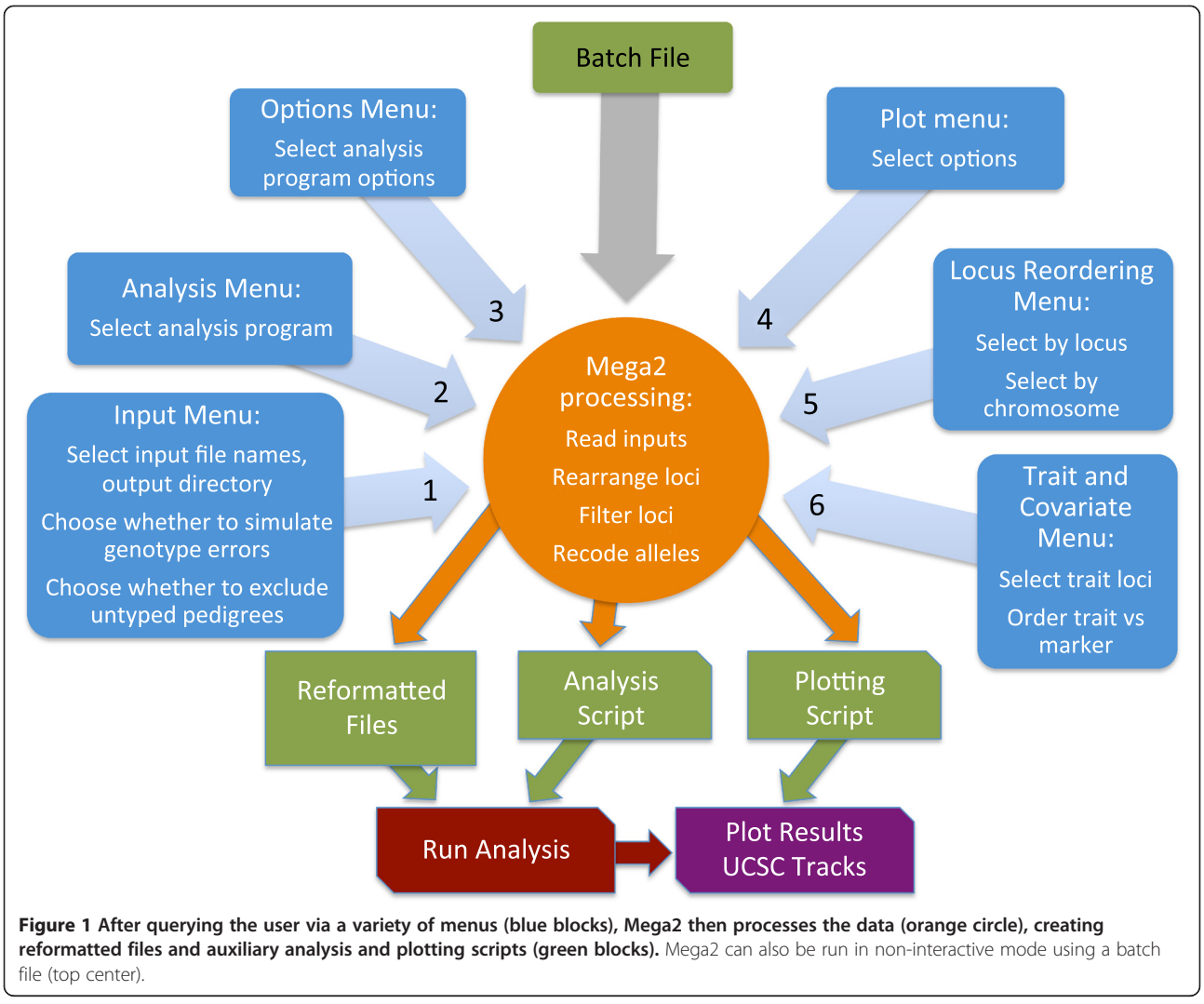

Figure 1 After querying the user via a variety of menus (blue blocks), Mega2 then processes the data (orange circle), creating reformatted files and auxiliary analysis and plotting scripts (green blocks). Mega2 can also be run in non-interactive mode using a batch file (top center).

as input and output, is important because it is a common way of compactly storing large scale data and provides a succinct way to efficiently get large scale data into and out of Mega2. Recently, seven new output formats have been added (see Table 1): (1) IQLS/Idcoefs [17-19] - a program for carrying out haplotype-based association tests while properly accounting for relatedness; (2) FBAT [20] – a program for carrying out family-based association tests; (3) Morgan [21] – a package capable of many analyses, with particular strengths in the area of Monte Carlo Markov Chain analyses of family data; (4) Beagle [22] – a package capable of many analyses, including haplotyping and association testing; (5) Eigenstrat [23,24] – a program for inferring and adjusting for population substructure from genome-wide marker data while testing for association; (6) Structure [25,26] – a program for investigating population structure and admixture using genome-wide marker data; and (7) PLINK/SEQ [27] – a package for analysis of data from large-scale sequencing projects.

The Mega2 distribution package has been updated to provide greater ease of installation and compatibility with many Unix environments. It contains added support for migration of legacy input data to our updated formats.

Discussion

In applied data analysis, a thorough analysis often requires the use of multiple different programs, many of which have their own precise input format requirements. Reformatting programs such as Mega2 can markedly accelerate analyses by providing accurate, quick, and error-free conversion routines. This need has been recognized in the area of population genetics, where several reformatting programs have been written [50-52], including one that converts to 52 different formats [51]. In the area of human genetics, limited reformatting options have been made available as part of larger database systems. For example, the GeneLink database system [53] initially exported into LINKAGE [11-13,54], GAS [55], or RelCheck [56,57]

Table 1 Mega2 currently supports 37 output targets seven new ones have been added since 2011

SimWalk2 format [28]	SOLAR format [4,5]	Mendel format [7]
Vintage MENDEL format [8]	Vitesse format [29]	SUP format [30,31]
ASPEX format	Linkage format [11-13] Pre-makeped format [11-13]	Cranefoot format [16]
GeneHunter-Plus format [32]	Testing loci for HWE	Mega2 annotated format
GeneHunter format [33,34]	Allegro format [35]	IQLS/ldcoefs format [17,36] (*added 6/11*)
Conversion to nuclear families	MLBQTL format [37]	PLINK format [14] (*binary added 1/13*)
SLINK format [30,31,38]	SAGE format [39]	FBAT format [20] (*added 1/13*)
SPLINK format [40]	Merlin/SimWalk2-NPL format	Morgan format [21] (*added 6/13*)
Homogeneity analyses	PREST format [2,3]	Beagle format [22,41-43] (*added 6/13*)
SIMULATE format [44]	PAP format [45,46]	Eigenstrat format [23,24] (*added 6/13*)
Genotype/phenotype/segregation summaries	Merlin format [47]	Structure format [25,26,48] (*added 6/13*)
Old SAGE format	Loki format [49]	PLINK/SEQ format [27] (*added 10/13*)

formats, while the Integrated Genotyping System [58] exported into several formats, including Merlin [47], GeneHunter [33,34], QTDT [59], and Transmit [60] formats. However, these database systems can be difficult to install and maintain. Other more stand-alone approaches to reformatting in this area include SIB-PAIR [61], by David Duffy, which is a command-line oriented program that can create locus and pedigree files in a variety of formats, such as FISHER [8], GAS [55], Genehunter [33,34], LINKAGE [11-13,54], LOKI [49], MENDEL [6], MERLIN [47], PAP [45,46] and SAGE [39]. SIB-PAIR appears to require very detailed line-by-line commands that would make it harder to use than Mega2 for most users. Another program is fcGENE [62] (available from SourceForge), which is focused on converting PLINK-format data for imputation (MaCH [63], IMPUTE [64], BEAGLE [22,41-43], BIMBAM [65]), and then converting the resulting imputed data into the following formats: PLINK [14], SNPTEST [64], HAPLOVIEW [66], EIGENSOFT [23,24], GenABEL [67], and VCF [15]. While fcGENE is fast and easy to use, it is currently limited (e.g., it does not accept VCF or LINKAGE format as input, it only supports a single dichotomous phenotype, it does not support selection by chromosome, etc.).

ALOHOMORA [68] provides an elegant interface for carrying out linkage analyses of Affymetrix 10K single nucleotide polymorphism (SNP) genotype data. This program actually uses Mega2 as its internal reformatting engine for some of its options.

PLINK [14] is an association analysis toolset that has a variety of data management and filtering options for handling large-scale SNP data. The main focus of PLINK is population-based unrelated samples, with some support for family-based association testing. PLINK only exports data in a few limited formats. We have used PLINK on our family data to carry out data cleaning, but then still needed Mega2 to reformat the data in order to carry out analyses using other external programs. In our experience, it is difficult to use PLINK on family data while maintaining the original pedigree structures upon output, as PLINK favors automatically filtering out individuals with low genotyping success rates (such as untyped founders).

From this brief survey of currently available data reformatting software, two things are immediately apparent: many researchers have recognized the need for providing one's data in many different formats; and Mega2, which is free, open source, and available on Unix, Windows, and Macintosh platforms, is well-positioned to continue to fill this need.

Conclusion

When carrying out quality control and statistical analyses for a genetic study of a human disease, one quickly discovers that data organization and analysis set-up is a critical, time-consuming, and extremely tedious task. Furthermore, one often needs to use several different analysis programs, each with its own idiosyncratic input format requirements. To meet these needs, we developed Mega2, taking the time to carefully understand the precise (sometimes poorly documented) requirements of each target format, implementing our data reformatting pipeline in tested and well-documented code. Mega2's tested and validated data conversion options expands the universe of possible analyses for the average researcher by removing the hurdle of having to tediously write, check, debug, and maintain their own conversion scripts.

Availability and requirements

Project name: Mega2
Project home page: https://watson.hgen.pitt.edu/register/
Operating systems: Linux, Macintosh OS X, Windows, Solaris
Programming language: C++
Other requirements: R, Perl, Python, awk, bash, and csh
License: GNU GPL v3
Any restrictions to use by non-academics: None.

Additional file

Additional file 1: A zipped archive containing the Mega2 version 4.7.1 distribution package; both source and binary executables are included.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RVB has been contributing to the Mega2 project as a programmer since 2010. CK contributed to the Mega2 project as a programmer from 2012 to 2014. NM was the primary programmer of Mega2 through 2010. DEW supervised the Mega2 project. All authors contributed to the writing of this article. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the National Institutes of Health grant R01 GM076667 (P.I. Daniel E. Weeks) and the University of Pittsburgh. We thank Lee Almasy, Mark Schroeder, and William P. Mulvihill for early contributions as programmers to our Mega2 project.

Author details
[1]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA. [2]Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA. [3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA.

Received: 18 July 2014 Accepted: 14 November 2014
Published online: 05 December 2014

References
1. Wigginton JE, Abecasis GR: **PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data.** *Bioinformatics* 2005, **21**(16):3445–3447.
2. Sun L, Wilder K, McPeek MS: **Enhanced pedigree error detection.** *Hum Hered* 2002, **54**(2):99–110.
3. McPeek MS, Sun L: **Statistical tests for detection of misspecified relationships by use of genome-screen data.** *Am J Hum Genet* 2000, **66**(3):1076–1094.
4. Almasy L, Blangero J: **Multipoint quantitative-trait linkage analysis in general pedigrees.** *Am J Hum Genet* 1998, **62**(5):1198–1211.
5. Blangero J, Almasy L: **Multipoint oligogenic linkage analysis of quantitative traits.** *Genet Epidemiol* 1997, **14**(6):959–964.
6. Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM: **Mendel: the Swiss army knife of genetic analysis programs.** *Bioinformatics* 2013, **29**(12):1568–1570.
7. Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E: **MENDEL version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets.** *Am J Hum Genet* 2001, **69**(Suppl):504.
8. Lange K, Weeks D, Boehnke M: **Programs for pedigree analysis: MENDEL, FISHER, and dGENE.** *Genet Epidemiol* 1988, **5**:471–472.
9. Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE: **Mega2: data-handling for facilitating genetic linkage and association analyses.** *Bioinformatics* 2005, **21**(10):2556–2557.
10. Mukhopadhyay 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* 1999, **65**:A436.
11. Lathrop GM, Lalouel J-M: **Easy calculations of lod scores and genetic risks on small computers.** *Am J Hum Genet* 1984, **36**:460–465.
12. Lathrop GM, Lalouel JM, Julier C, Ott J: **Strategies for multilocus linkage analysis in humans.** *Proc Natl Acad Sci U S A* 1984, **81**:3443–3446.
13. Lathrop GM, Lalouel JM: **Efficient computations in multilocus linkage analysis.** *Am J Hum Genet* 1988, **42**:498–505.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559–575.
15. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27**(15):2156–2158.
16. Makinen VP, Parkkonen M, Wessman M, Groop PH, Kanninen T, Kaski K: **High-throughput pedigree drawing.** *Eur J Hum Genet* 2005, **13**(8):987–989.
17. Wang Z, McPeek MS: **An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related Individuals.** *J Am Stat Assoc* 2009, **104**(487):1251–1260.
18. Abney MA, Ober C, McPeek MS: **Homozygosity mapping of quantitative trait loci in complex inbred pedigrees.** *Am J Hum Genet* 2000, **67**(Suppl 2):327.
19. Wang Z, McPeek MS: **ATRIUM: testing untyped SNPs in case-control association studies with related individuals.** *Am J Hum Genet* 2009, **85**(5):667–678.
20. Laird NM, Horvath S, Xu X: **Implementing a unified approach to family-based tests of association.** *Genet Epidemiol* 2000, **19**(Suppl 1):S36–42.
21. Thompson EA: *Statistical inference from genetic data on pedigrees, vol. 6.* Beechwood, OH: Institute of Mathematical Sciences and the American Statistical Association; 2000.
22. Browning BL, Browning SR: **Efficient multilocus association testing for whole genome association studies using localized haplotype clustering.** *Genet Epidemiol* 2007, **31**(5):365–375.
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**(8):904–909.
24. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2**(12):e190.
25. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**(2):945–959.
26. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.** *Genetics* 2003, **164**(4):1567–1587.
27. PLINK/SEQ: *A library for the analysis of genetic variation data*; [http://atgu.mgh.harvard.edu/plinkseq/]
28. Sobel E, Lange K: **Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics.** *Am J Hum Genet* 1996, **58**(6):1323–1337.
29. 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.
30. Lemire M: **SUP: an extension to SLINK to allow a larger number of marker loci to be simulated in pedigrees conditional on trait values.** *BMC Genet* 2006, **7**:40.
31. Schäffer AA, Lemire M, Ott J, Lathrop GM, Weeks DE: **Coordinated conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrees.** *Hum Hered* 2011, **71**(2):126–134.
32. Kong A, Cox NJ: **Allele-sharing models: LOD scores and accurate linkage tests.** *Am J Hum Genet* 1997, **61**(5):1179–1188.
33. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: **Parametric and nonparametric linkage analysis: a unified multipoint approach.** *Am J Hum Genet* 1996, **58**:1347–1363.
34. Kruglyak L, Lander ES: **Faster multipoint linkage analysis using Fourier transforms.** *J Comput Biol* 1998, **5**(1):1–7.
35. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: **Allegro, a new computer program for multipoint linkage analysis.** *Nat Genet* 2000, **25**(1):12–13.
36. Abney M, McPeek MS, Ober C: **Estimation of variance components of quantitative traits in inbred populations.** *Am J Hum Genet* 2000, **66**(2):629–650.
37. Alcais A, Abel L: **Maximum-Likelihood-Binomial method for genetic model-free linkage analysis of quantitative traits in sibships.** *Genet Epidemiol* 1999, **17**(2):102–117.
38. Weeks DE, Ott J, Lathrop GM: **SLINK: a general simulation program for linkage analysis.** *Am J Hum Genet* 1990, **47**(3):A204.
39. S.A.G.E: *Statistical Analysis for Genetic Epidemiology*; [http://darwin.cwru.edu/sage/]
40. Holmans P: **Asymptotic properties of affected-sib-pair linkage analysis.** *Am J Hum Genet* 1993, **52**(2):362–374.
41. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**(2):210–223.
42. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81**(5):1084–1097.

43. Browning SR, Briley JD, Briley LP, Chandra G, Charnecki JH, Ehm MG, Johansson KA, Jones BJ, Karter AJ, Yarnall DP, Wagner MJ: **Case-control single-marker and haplotypic association analysis of pedigree data.** *Genet Epidemiol* 2005, **28**(2):110–122.
44. Terwilliger JD, Speer M, Ott J: **Chromosome-based method for rapid computer simulation in human genetic linkage analysis.** *Genet Epidemiol* 1993, **10**(4):217–224.
45. Hasstedt SJ: **jPAP: Document-driven software for genetic analysis.** *Genet Epidemiol* 2005, **29:**255.
46. PAP: *Pedigree Analysis Software*; [http://hasstedt.genetics.utah.edu/]
47. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin–rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30**(1):97–101.
48. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: dominant markers and null alleles.** *Mol Ecol Notes* 2007, **7**(4):574–578.
49. Heath SC: **Markov chain Monte Carlo segregation and linkage analysis for oligogenic models.** *Am J Hum Genet* 1997, **61**(3):748–760.
50. Manoukis NC: **FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis.** *Mol Ecol Notes* 2007, **7**(4):592–593.
51. Coombs JA, Letcher BH, Nislow KH: **CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs.** *Mol Ecol Resour* 2008, **8**(3):578–580.
52. Glaubitz JC: **CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages.** *Mol Ecol Notes* 2004, **4**(2):309–310.
53. Gillanders EM, Masiello A, Gildea D, Umayam L, Duggal P, Jones MP, Klein AP, Freas-Lutz D, Ibay G, Trout K, Wolfsberg TG, Trent JM, Bailey-Wilson JE, Baxevanis AD: **GeneLink: a database to facilitate genetic studies of complex traits.** *BMC Genomics* 2004, **5**(1):81.
54. Lathrop GM, Lalouel JM, Julier C, Ott J: **Multilocus linkage analysis in humans: detection of linkage and estimation of recombination.** *Am J Hum Genet* 1985, **37**(3):482–498.
55. GAS: *Genetic Analysis System*; [http://users.ox.ac.uk/~ayoung/gas.html]
56. Epstein MP, Duren WL, Boehnke M: **Improved inference of relationship for pairs of individuals.** *Am J Hum Genet* 2000, **67**(5):1219–1231.
57. Boehnke M, Cox NJ: **Accurate inference of relationships in sib-pair linkage studies.** *Am J Hum Genet* 1997, **61**(2):423–429.
58. Fiddy S, Cattermole D, Xie D, Duan XY, Mott R: **An integrated system for genetic analysis.** *BMC Bioinformatics* 2006, **7:**210.
59. Abecasis GR, Cardon LR, Cookson WO: **A general test of association for quantitative traits in nuclear families.** *Am J Hum Genet* 2000, **66**(1):279–292.
60. Clayton D: **A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission.** *Am J Hum Genet* 1999, **65**(4):1170–1177.
61. *SIB-PAIR*; [http://genepi.qimr.edu.au/staff/davidD/]
62. fcGENE: *Genotype format converter*; [http://sourceforge.net/projects/fcgene/]
63. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: **MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.** *Genet Epidemiol* 2010, **34**(8):816–834.
64. Marchini J, Howie B: **Genotype imputation for genome-wide association studies.** *Nat Rev Genet* 2010, **11**(7):499–511.
65. Servin B, Stephens M: **Imputation-based analysis of association studies: candidate regions and quantitative traits.** *PLoS Genet* 2007, **3**(7):e114.
66. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**(2):263–265.
67. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, **23**(10):1294–1296.
68. Ruschendorf F, Nurnberg P: **ALOHOMORA: a tool for linkage analysis using 10K SNP array data.** *Bioinformatics* 2005, **21**(9):2123–2125.

doi:10.1186/s13029-014-0026-y
Cite this article as: Baron *et al.*: Mega2: validated data-reformatting for linkage and association analyses. *Source Code for Biology and Medicine* 2014 **9**:26.