BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2128–2129
doi:10.1093/bioinformatics/bti282

Genetics and population analysis

PowerMarker: an integrated analysis environment for genetic marker analysis

Kejun Liu† and Spencer V. Muse*

Bioinformatics Research Center, Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566, USA

Received on May 12, 2004; revised on January 17, 2005; accepted on January 18, 2005
Advance Access publication February 10, 2005

ABSTRACT

Summary: PowerMarker delivers a data-driven, integrated analysis environment (IAE) for genetic data. The IAE integrates data management, analysis and visualization in a user-friendly graphical user interface. It accelerates the analysis lifecycle and enables users to maintain data integrity throughout the process. An ever-growing list of more than 50 different statistical analyses for genetic markers has been implemented in PowerMarker.

Availability: www.powermarker.net

Contact: powermarker@hotmail.com

INTRODUCTION

Fundamental and applied population genetics, quantitative genetics and human genetics depend heavily on the availability of genetic markers. Markers are used by population geneticists to investigate the origin, genetic diversity and population structure of alleles, by evolutionists to describe genetic relationships among species or populations and by geneticists to study linkage disequilibrium (LD) within or between genes. Markers are now widely used in the search for genes affecting human diseases by identifying statistical associations between the genetic markers and the traits of interest. Traditionally, these tasks have been performed by a variety of separate tools (Felsenstein, 1993; Lewis and Zaykin, 2001, Free program distributed over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html), each of which having its own specific input and output formats. Such tools provide geneticists the necessary functionality to analyze their data. However, many of these tools, implemented as standalone programs, are not especially user-friendly and require the users to spend considerable time on data preparation and/or result parsing. As far as the authors can tell, there is no publicly available package for genetic marker analysis that integrates the data management, statistical analysis and visualization aspects of genetic data analysis through a single user-friendly graphical interface. This note describes such an application that allows scientists to perform genetic marker data analysis in an integrated analysis environment (IAE).

*To whom correspondence should be addressed.

†Present address: GlaxoSmithKline, PO Box 13398, Five Moore Drive, Research Triangle Park, NC, USA

OVERVIEW

PowerMarker includes a powerful graphical interface (http://www.powermarker.net). The user interface allows the user to manage projects and a variety of data objects, to perform over 50 different analyses in more than 20 modules, and to manipulate and view all data objects. A 'project' in PowerMarker consists of data objects (datasets, tables, etc.) and folders. Data objects in a project are organized by different object types. Users can create folders and give shortcuts to data objects in a folder by simple drag and drop operations. Upon input, all data are serialized as binary formats to reduce storage demands and improve computational efficiency. Perhaps more importantly, serialized data is not human readable and is not subject to casual editing, a feature useful for error reduction. Of course, PowerMarker can produce human readable datasets from the binary forms.

PowerMarker handles a variety of marker data, including both haplotypes and diplotypes. The gametic phase for the diplotype data can be known or unknown. Examples of marker data include microsatellite data, single nucleotide polymorphism (SNP) data, and Restriction Fragment Length Polymorphism (RFLP) data. PowerMarker does not require a specific input format such as NEXUS format. Instead, PowerMarker supports table-like format directly. When importing a dataset, the user can choose one of the two possible modes: the dataset wizard will guide the user step-by-step through the process of data importing or the batch importer can simultaneously import multiple datasets with the same format. PowerMarker can export datasets in a variety of formats.

METHODS IMPLEMENTED IN POWERMARKER

PowerMarker computes several summary statistics for each marker locus, including allele number, missing proportion, heterozygosity, gene diversity, polymorphism information content (PIC) and stepwise patterns for microsatellite data. Variances and confidence intervals of these statistics are estimated by non-parametric bootstrapping across different loci. Allele frequencies and genotype frequencies are estimated by simple counting, while haplotype frequencies are estimated by the widely used EM algorithm (Excoffier and Slatkin, 1995). The variances and confidence intervals of these frequencies are estimated by bootstrapping across individuals. Haplotype estimation in PowerMarker is highly optimized for SNP data by taking advantage of the binary feature of these markers. We also

have efficient EM algorithms for trio families (Rohde and Fuerst, 2001).

PowerMarker implements all common methods for testing Hardy–Weinberg equilibrium and linkage equilibrium, including χ^2 tests, likelihood ratio tests and exact tests. Most common measures of LD are calculated by PowerMarker, including D' and r^2. PowerMarker also creates several LD-distance plots directly in Microsoft Excel. The LD matrix constructed in PowerMarker can be viewed internally by its 2D viewer and exported as a graphics file for further editing.

Differentiation among populations is often summarized using F-statistics. PowerMarker performs four different types of F-statistics analysis. Several data selection modules are included in PowerMarker for experimental design purposes. The goal of line selection is to choose a core set of lines with maximal gene diversity from a larger germplasm collection. We implemented a flexible simulated annealing algorithm to do the combinatorial optimization (K.Liu, Y.Xiang and S.V.Muse, submitted for publication). A unique feature of the algorithm is that general constraints can be incorporated in the algorithm. The idea underlying marker selection is to choose relatively uncorrelated markers from a larger pool considering linkage disequilibrium between markers. Phylogenetic analysis in PowerMarker covers four modules: the frequency module computing allele frequencies, the distance module computing 19 different distances based on allele frequencies or repeat patterns of microsatellite markers, the tree module computing trees from distances and the bootstrap module that generates a list of trees by bootstrapping across markers.

PowerMarker offers three methods for testing an association between a single marker and the affected status of individuals (must be binary): the allele case-control test, genotypic case-control test and the multiallelic trend test. An F-test is provided for quantitative traits. Haplotype trend regression (Zaykin *et al*., 2002) can be applied to both quantitative traits and binary traits. The coalescence simulation with hotspot recombination model of Posada and Wiuf (2003) is implemented in PowerMarker. With no hotspot defined, the simulation becomes the classical coalescence model with homogeneous recombination (Hudson, 1983; Hudson and Kaplan, 1990). PowerMarker's SNP identification tool identifies SNPs from sequence data. A variety of user-settable options are available for this tool. PowerMarker also includes modules for Mantel tests and contingency table analyses.

For details on using the software and explanations of the underlying algorithms, we refer readers to the manual and the references listed there. Currently PowerMarker does not implement any linkage analyses, but this point will be addressed in the future version.

SPECIAL FEATURES

PowerMarker offers several special features for visualization and data analysis:

- Two-dimensional (2D) plots and triangle plots: 2D plots are used extensively for visualizing linkage disequilibria results. The 2D plot module in PowerMarker provides a powerful editor for visualizing two-way tables. The resulting plot can be saved as a Windows Meta File (WMF) for further editing. Triangle plots are useful for characterizing population structure.
- Excel integration: Datasets and tables in PowerMarker can be opened in Excel by double-clicking in the internal viewers. PowerMarker also directly draws triangle plots in Excel.
- Multithreading batch system: all of the analyses and data manipulations in PowerMarker can support multiple datasets in a graphical user interface. Also, each analysis runs in its own thread, so it is possible to pause/resume or cancel an analysis without affecting the graphical interface or other running analyses. For computers with multiple CPUs, the multithreading system makes it possible to allocate multiple CPUs to the program.

IMPLEMENTATION

The PowerMarker package was written in Visual C# and runs under the Microsoft .NET framework. The numerical library in PowerMarker was written in Visual C++. Excel integration was implemented through DCOM. The software does not put a limitation on the sample size or marker number of the dataset (some analyses, such as haplotype frequency estimation, will be subject to the limitation of the size of computer memory). The authors are frequently updating the software to support new analyses.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the John Doebley and Ed Buckler labs for their significant and continuing support. Input from Bruce Weir and Elizabeth Thompson is recognized and appreciated. The development of PowerMarker was partly supported by NSF grants DBI-0096033 and DEB-9996118.

REFERENCES

Excoffier,L. and Slatkin,M. (1995) Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.

Felsenstein,J. (1993) PHYLIP (phylogeny Inference Package), version 3.5c. Department of Genetics, University of Washington, Seattle.

Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.*, **23**, 183–201.

Hudson,R.R. and Kaplan,N. (1990) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Lewis,P.O. and Zaykin,D. (2001) Genetic data analysis: computer program for the analysis of allelic data. Version 1.0.

Posada,D. and Wiuf,C. (2003) Simulating haplotype blocks in the human genome. *Bioinformatics*, **19**, 289–290.

Rohde,K. and Fuerst,R. (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum. Mutat.*, **17**, 289–295.

Zaykin,D.V., Westfall,P.H., Young,S.S., Karnoub,M.C., Wagner,M.J. and Ehm,M.G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79–91.