Integrated analysis of genetic data with R

Jing Hua Zhao[1*] *and Qihua Tan*[2]

[1] MRC Epidemiology Unit, Strangeways Research Laboratory, Wort's Causeway, Cambridge, CB1 8RN, UK
[2] Odense University Hospital, KKA, Department of Clinical Biochemistry and Genetics, Sdr. Boulevard 29, DK-5000, Odense C, Denmark, Tel: +45 65412822; Fax: +45 65411911; qihua.tan@ouh.fyns-amt.dk
**Correspondence to*: Tel: +44 1223 741398; Fax: +44 1223 740050; E-mail: jhz22@medschl.cam.ac.uk

Date Received: 4th July 2005

Abstract

Genetic data are now widely available. There is, however, an apparent lack of concerted effort to produce software systems for statistical analysis of genetic data compared with other fields of statistics. It is often a tremendous task for end-users to tailor them for particular data, especially when genetic data are analysed in conjunction with a large number of covariates. Here, R (http://www.r-project.org), a free, flexible and platform-independent environment for statistical modelling and graphics is explored as an integrated system for genetic data analysis. An overview of some packages currently available for analysis of genetic data is given. This is followed by examples of package development and practical applications. With clear advantages in data management, graphics, statistical analysis, programming, internet capability and use of available codes, it is a feasible, although challenging, task to develop it into an integrated platform for genetic analysis; this will require the joint efforts of many researchers.

Keywords: *linkage and association analysis, complex traits, software*

Introduction

With the success of genome projects in human and other species, vast quantities of genetic data are now available and increasingly used. These include the HapMap (http://www.hapmap.org) and the BioBank (eg UK BioBank, http://www.ukbiobank.ac.uk) projects; others are envisaged.[1] They generate large datasets, which are used for localisation of disease-predisposing genes, for drug discovery and for better understanding of human population history and interaction with the environment.

Meanwhile, these data and their increasing use pose immense challenges for statisticians and have provoked a bewildering array of new algorithms and relevant software (for example, in phasing algorithms[2,3]). There is an apparent lack of coordination of such endeavours, however, compared with other fields of statistics, where appropriate tools are well established. For human genetics, the focus of research has been on the genetic dissection of complex traits such as schizophrenia, diabetes and cardiovascular diseases.
The research paradigms and tools largely fall into several categories; namely, segregation analysis, linkage analysis (including allele-sharing methods), association studies and experimental crosses mapping polygenic traits, mapping of quantitative trait loci (QTLs).[4] This has been further shifted to genetic association studies exploring genomic structure and incorporating more information regarding human population history, microarray analysis using gene expression data and proteomics, among others. The vast genomic data narrow the gap between the original definitions of genetic mapping and sequence analysis in the human genome project, followed by a similar trend for analytical tools. These analytical tools now appear awkward and require updating.

There are hundreds of programs and utilities for linkage and association analysis. Some of them are described here. Since most of them are listed in the Rockefeller University (http://linkage.rockefeller.edu) and UK Human Genome Mapping Project Resource Centre (http://www.hgmp.mrc.ac.uk). The full names and references for these software programs are not given here. This paper is placed in the context of previous reviews on linkage analysis[5] and haplotype phase inference.[2] It is notable that Salem *et al's*[3] survey contained a total of 43 software programs for phasing and association analysis of unrelated individuals. It would have been much more if topics such as data on experimental design using animals, phylogenetic analysis and microarray data analysis had been included.

Computer programming for linkage analysis began with the first Fortran program, LIPED, developed by Ott.[6,7] In the 1980s, the celebrated book *Methods in Genetic Epidemiology*[8] described a variety of computer programs, including PATHMIX for path analysis of nuclear family data, POINTER for complex segregation analysis and LIPED. The Pascal program LINKAGE was written in the early 1980s and included a number of subprograms, ILINK, MLINK and LINKMAP, which in turn had their counterpart adapted for three-generation Centre d'Etude du Polymorphisme

Humain families. These programs are still widely used, but require intense training. Based on these original programs, a number of other programs have been written — for example, SLINK, FASTSLINK, FASTLINK, MFLINK, FASTMAP, ERPA, ESPA and a variety of linkage utility programs. Added to the analysts' learning set are packages such as MENDEL, SIMLINK, SIMWALK, VITESSE, PAP, SAGE, SOLAR, SUPERLINK, SPLINK and ASPEX. Unfortunately, these are not exclusive; for example, a number of programs based on the Lander-Green-Kruglyak algorithm have been developed; for example, MAPMAKER, GENEHUNTER, GENEHUNTER-PLUS, GENEHUNTER-IMPRINTING, GENEHUNTER-TWOLOCUS, GENEHUNTER-SAD, ALLEGRO, MERLIN. There are also programs based on Bayesian methods such as MORGAN. Furthermore, efforts have been devoted to developing tools to facilitate analysis; for example, GLUE, QUICKLINK and easyLINKAGE. Popular programs for phasing and association analysis included ARLEQUIN, PHASE, EHPLUS, SNPHAP, PLEM and CHAPLAN for unrelated individuals and QTDT, FBAT, TRANSMIT and UNPHASED for family-based association tests. There are also Bayesian counterparts such as HAPLOTYPER and BLADE.

Some features of these programs are worthy of note. First, they were written in many computer languages, ranging from C, C++, Fortran, Pascal, Java and Perl to Stata, SAS and S-PLUS, some of which are available in compiled form and are tested under specific computer systems. Secondly, they require data in specific formats, often from the programmers' own perspective and not conforming to any standard, and it is often rather cumbersome to reuse output from these programs. Some include primitive parsing and a few have graphical capability. Thirdly, in the analysis of data from a large project, it is often necessary to write some customised utilities for these programs. The batch of skills required for the different languages and tools largely needs a profession of computing or an applied field. These often lead to redundant work, poor maintenance and lack of validity checks. Consequently, it is difficult for practical data analysts to keep track of so many software programs and thus many smaller programs are sometimes 'lost', even though they would be very useful if only people knew about them.

The features of good software systems for genetic data analysis have been described[9] and were reiterated in the recent Genetic Analysis Workshop (http://www.gaworkshop.org), where some software proved to be inadequate for datasets from both real study and simulation. To a large extent, the authors believe that this is due to the lack of a general but satisfactory platform for statistical geneticists. Excellent theoretical work often does not have a good companion program. While there is always a motivation to provide one, the effort of development is often too great. The ideal development platform should run across computer systems and have facilities for data management, graphics, established algorithms and clear documentation, provide a graphical user-interface (GUI) and accept batch jobs. The language should be powerful and flexible, but easy enough to track errors and modify or extend the source codes. Furthermore, it is essential to be able to retrieve and send information from the internet, given that large genetic data and programs are publicly available. Finally, as much of the code for numerical analysis and other routines has been available for decades, typically in Pascal, Fortran, C/C++, it should ideally be possible to re-use this.

The above features would be impossible to achieve by single programmer(s) or group(s); however, with the recent development of general computing, such a platform now exists. Note how these features are reminiscent of the open source initiatives led by the Linux operating system. In the following sections, the authors first describe features of R through a brief introduction, and then give a survey of packages to illustrate the range of tools available. This is followed by an exposition through example packages. They also provide examples and comparisons with other platforms. They suggest that R could potentially serve as an integrated platform for genetic data analysis.

A brief overview of R

According to the comprehensive R archive network (CRAN),

R is "GNU S", a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc' (http://cran.r-project.org).

A brief history is contained in the frequently asked questions (R-FAQ) at CRAN:

The name is partly based on the (first) names of the first two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs language "S". S is a very high level language and an environment for data analysis and graphics. In 1998, the Association for Computing Machinery (ACM) presented its Software System Award to John M. Chambers, the principal designer of S, for the S system, which has forever altered the way people analyze, visualize, and manipulate data... S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers.

The master site for CRAN is maintained in Austria and is mirrored by other sites worldwide. The R system is greatly enhanced by a variety of tools with excellent documentation. These tools are organised as base and contributed packages, which now number well over 500. Similarly to S-PLUS, an R package is a collection of object(s), dataset(s) or function(s) for specific tasks. As of the latest version (2.1.1), the R distribution comes with the following packages: base R functions (*base*); base R datasets (*datasets*); formally defined methods and classes for R objects and programming tools (*methods*); devices and functions for graphics (*grid, grDevices*,

graphics); interface and language bindings to Tcl/Tk GUI elements (*tcltk*); tools for package development administration (*tools*); utilities (*utils*); and R statistical functions (*stats, stats4, splines*). The recommended packages include bootstrap (*boot*); cluster analysis (*cluster*); interface to other statistical packages (*foreign*); lattice graphics (*lattice*); linear models and smoothing (*mgcv, nlme, KernSmooth*); recursive partitioning (*rpart*); survival analysis (*survival*); and functions and datasets to support the book *Modern Applied Statistics* with S[10] (*VR*). By default, these packages are installed with the R system. By contrast, contributed packages are from other users and require *install.packages()* for installation and *library()* command to load.

The R system is available for most computer systems, including Unix, Linux, Windows and MacOS X. With the R system is an object-orientated programming language, a powerful tool for organising the representation of information (classes) and the actions that are applied to these representations (methods). It now supports the S4 class system,[11] which is distinguished from the S3 class system[12] and allows for object-orientated programming within an interactive environment, consistent validity check and multiple method dispatches. In addition, R has a flexible graphical facility and is able to read data in a number of formats, including dBase, Stata, SPSS and SAS. It also includes a linear algebra package (LAPACK, http://www.netlib.org/lapack/). As R was developed using the model of S, from an institution where the C/C++ language was born, it is native to C/C++ and Fortran programs. Furthermore, it can be run both through GUI and in batch mode, which allows new and experienced users to customise it to their own needs. TCL/Tk is now part of the system, which may be used to create a user-defined GUI. There are also packages which provide interface to common gateway interface (CGI) and generate HTML/XML outputs. A closely related project is Omega, '...a joint project with the goal of providing a variety of open-source software for statistical applications' (http://www.omegahat.org), which aims to provide facilities to communicate between R and other applications such as Matlab, Perl and Python. The packages RMySQL and RODBC are useful for connecting the MySQL database system and Open DataBase Connectivity (ODBC).

More information about R, including documentation and recommended reading, is available from CRAN.

A list of R packages for genetic data analysis

This section describes some packages for genetic data analysis according to their package descriptions in CRAN. They fall into several categories: data manipulation (*genetics*); phylogenetic analysis (*PHYLOGR, ape*); association analysis of population data including population structure (*biodem*, *genetics*, *hapassoc*, *haplo.score*, *haplo.stats*, *hierfstat*, *hwde*, *ldDesign*, *LDheatmap*, *Malmig*, *popgen*, *R/gap*, *rmetasim*); family data (*tdthap*); and QTL for experimental design (*bim*, *bqtl*, *happy*, *qtlDesign*, *R/qtl*). Others (*BradleyTerry*, *epitools*, *evd*, *gllm*, *locfdr*, *rmeta*, *vcd*) are fairly general and are not limited to analysis of genetic data. There are a large number of packages for microarray analysis, as described below.

ape. Analyses of Phylogenetics and Evolution: provides functions for reading and plotting phylogenetic trees in parenthetic format (standard Newick format), analyses of comparative data in a phylogenetic framework, analyses of diversification and macroevolution, computing distances from allelic and nucleotide data, reading nucleotide sequences from GenBank via the internet, and several tools such as Mantel's test, computation of minimum spanning tree or the population parameter theta based on various approaches.

bim. Bayesian interval mapping diagnostics: functions to interpret QTLCart and Bmapqtl samples.[13]

biodem. Biodemography functions.

bqtl. QTL mapping toolkit for inbred crosses and recombinant inbred lines. Includes maximum likelihood and Bayesian tools.

genetics. Classes and methods for handling genetic data. Includes classes to represent genotypes and haplotypes at single markers up to multiple markers on multiple chromosomes. Functions include allele frequencies, flagging homo/heterozygotes, flagging carriers of certain alleles, estimating and testing for Hardy–Weinberg disequilibrium, estimating and testing for linkage disequilibrium.

hapassoc. A package used for likelihood inference of trait associations with haplotypes and other covariates in generalised linear models. The functions accommodate uncertain haplotype phase and can handle missing genotypes at some SNPs.[14]

haplo.score. A suite of routines that can be used to compute score statistics to test associations between haplotypes and a wide variety of traits, including binary, ordinal, quantitative and Poisson.[15] These methods assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic markers). The methods provide several different global and haplotype-specific tests for association, as well as provide adjustment for non-genetic covariates and computation of simulation *p*-values (which may be needed for sparse data).

haplo.stats. A suite of S-PLUS/R routines for the analysis of indirectly measured haplotypes.[16] The statistical methods assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic

markers). The genetic markers are assumed to be co-dominant (ie one-to-one correspondence between their genotypes and their phenotypes), and the measurements of genetic markers are referred to as genotypes. The main functions in haplo.stats are: *haplo.em*, *haplo.glm* and *haplo.score*. The *haplo.score* function is an extension of an earlier function in the *haplo.score* package.

hierfstat. Estimation of hierarchical F-statistics from haploid or diploid genetic data with any numbers of levels in the hierarchy, and tests for the significance of each F and variance components.[17]

hwde. fits models for genotypic disequilibria, as described by Weir and Wilson[18] and Huttley and Wilson.[19] Contrast terms are available which account for the first-order interactions between loci.

kinship. A package that contains several functions. 1. coxme (general mixed-effects Cox models, kinship): routines to create and manipulate *n* by *n* matrices that describe the genetic relationships between *n* persons. 2. pedigree: creates and plots pedigrees. 3. bdsmatrix: a class of objects for sparse block-diagonal matrices (which is how kinship matrices are stored). 4. gchol: generalised Cholesky decompositions.

ldDesign. A package for design of experiments for association studies for detection of linkage disequilibrium. Uses an existing deterministic power calculation for detection of linkage disequilibrium between a biallelic QTL and a biallelic marker, together with the Spiegelhalter and Smith–Bayes factor to generate designs with power to detect effects with a given Bayes factor.[20]

LDheatmap. A package to create a heat map (a false colour image with a dendrogram added to the left side and to the top) of linkage disequilibrium involving SNPs, using both r and D'.

Malmig. Malecot migration model functions.

PHYLOGR. Manipulation and analysis of phylogenetically-simulated datasets (as obtained from PDSIMUL in package PDAP) and phylogenetically-based analyses using GLS.

qtlDesign. Tools for the design of QTL experiments.[21]

R/gap. An integrated package for genetic data analysis of both population and family data. It contains functions for sample size calculations of both population- and family-based designs, probability of familial disease aggregation, kinship calculation, some statistics in linkage analysis and association analysis involving one or more genetic markers, including haplotype analysis. The functions included are: *hwe*, *hwe.hardy* for Hardy–Weinberg equilibria involving SNPs and highly polymorphic microsatellite markers; *s2k*, *gcontrol* for single-locus association analysis of polymorphic markers and genomic control;[22,23] *genecounting*; *gcp* for haplotype analysis of all chromosomes and missing data[24] and permutation tests; *tbyt*, *kbyl* for linkage disequilibrium statistics for SNPs and multiallelic markers; *htr*, *hap.score* for extracting haplotype information for haplotype trend regression analysis and regression incorporating covariates based on conditional regression, as implemented in the haplo.score package.[15] For family data, it includes family plotting through graphviz (*pedtodot*), exact probability of familial clustering disease (*pfc* and *pfc.sim*),[25] kinship calculation, involves genetic index of familiality (*gif*) and a simple kinship calculation (*kin.morgan*). Currently, it is bundled with an experimental version of POINTER and PATHMIX.[8]

rmetasim. An interface between R and the metasim simulation engine.[26] Facilitates the use of the metasim engine to build and run individual-based population genetics simulations.

R/qtl. Analysis of experimental crosses to identify QTLs.[27]

The following packages are not available from CRAN, but conform to the R standard:

happy. an R interface into the C package HAPPY for fine-mapping QTL in heterogeneous stocks,[28] which is an advanced intercross between (usually eight) founder inbred strains of mice suitable for fine-mapping QTL. The *happy* package is an extension of the original C program happy; it uses the C code to compute the probability of descent from each of the founders, at each locus position, but *happy* allows a much richer range of models to be fit to the data.

tdthap. Transmission/disequilibrium tests (TDT) for extended haplotypes, according to Clayton and Jones.[29]

popgen. A package which implements a variety of statistical and population genetic methodology.[30]

An *ld2* function for two-locus log-linear models is available from the *gllm*, routines for log-linear models of incomplete contingency tables,[31] including some latent class models via expectation maximisation (EM) and Fisher scoring approaches. Basic tools for applied epidemiology are implemented in the general-purpose package *epitools*, and the *visualizing categorical data*[32] (*vcd*) package Woolf's test includes for homogeneity on $2 \times 2 \times k$- tables over strata (ie if the log odds ratios are the same in all strata). The *locfdr* package is for computation of local false discovery rates.[33] The *rmeta* package contains many functions for meta-analysis which would be appropriate for the genetic analysis setting, while the *BradleyTerry* package can be used for TDT analysis. A potentially useful package for genome-wide association analysis is

evd, implementing extreme value distribution. R functions associated with specific papers include *link/tdt*,[34,35] *EHP*,[36] *tdtexact*[37] and *htpower/Nstage*.[38,39] A number of R programs, including those for methods of genomic controls, are available from the University of Pittsburg computational genetics lab (http://wpicr.wpic.pitt.edu/wpiccompgen/). They use the familiar format of input/output files, but are somewhat informal compared with many packages on CRAN.

Many packages for microarray data analysis are available from CRAN and the Bioconductor project (http://www.bioconductor.org); for example, *affy* for Affymetrix, *marray* and *arrayMagic* for cDNA data processing and packages for extracting signals from the scanner (*Spot*), for gene annotation and delineating biological pathways (*annotate*). Unlike genotype data, gene expression data–after data pre-processing including normalisation–can generally be analysed using the recommended packages installed with R for standard statistical analysis. Bioconductor additionally provides packages for adjusting for multiple testing (*multtest*), which is a typical issue in analysing high-dimensional microarray data. Taking advantage of the extensive graphical abilities of R, the package *geneplotter* allows users to associate microarray expression data with chromosomal location and to visualise their data using whole genome or single chromosome plots. The package *Rgraphviz* can be used for laying out biological pathways.

The numerous packages available may appear daunting, but a recent feature of R is the so-called CRAN task views, which allow users to browse packages by topic and provide tools to automatically install all packages for special areas of interest. A version for genetic analysis has been developed by Gregor Gorjanc (http://www.bfro.uni-lj.si/MR/ggorjan/software/R/Genetics.html) and will be available soon.

Example applications

In this section, some examples are given to illustrate the development and use of the R packages described above.

Example 1. Haplotype frequency estimation including haplotype association with case-control data. A number of computer programs have been written by one of the authors (J.H.Z.) for this purpose: 2LD,[40] EH + ,[41] fastEH + ,[42] GENECOUNTING and HAP.[24,43] They have now been integrated into functions available from *R/gap*, so that haplotype frequencies can be estimated using the EM algorithm,[44] including data on Chromosome X, to be served as input for *tbyt* or *kbyl* to obtain linkage disequilibrium measures such as D' and r^2 and linkage disequilibrium heat map. Instead of calling the executable files with utilities such as LDSHELL,[45] a simple loop is sufficient to run a sliding windows analysis and for estimation using data from several populations. Furthermore, haplotype assignment can be read into R for haplotype trend regression[46] of cross-sectional or longitudinal data. In addition, some well-known datasets[47,48] can be stored in compact form with detailed documentation and retrieved when needed.

Example 2. A collaborative study on genetics of alcoholism (COGA) data from the Genetic Analysis Workshop 14 (GAW14). The microsatellite markers are given in fixed ASCII format. In previous analysis,[49] C utility programs had to be written to read the marker data in allele size. Now, one can use *read.fortran* to read such formatted data. One can also use the *genetics* package to test for Hardy–Weinberg Equilibrium, and pedigree diagrams can be drawn all in one go for the 143 pedigrees involved (see http://www.ucl.ac.uk/~rmjdjhz/r-progs.htm). The use of the *kinship* package for the mixed-effects Cox model of alcoholism in extended pedigrees, including family relationship and with microsatellite markers, has been reported.[50]

Example 3. Log-linear models for genotype data. The R package, *hwde*, has provided an example from Huttley and Wilson;[19] see the detailed information given in the package vignette.

Example 4. Bayesian analysis of population data. This can be achieved with the R package *mcmc* by Charles Geyer. One can also use the more familiar Windows packages *WinBUGS*, via the package *R2WinBUGS*. It is easier to set up than the original *WinBUGS* and the convergence of Markov chain Monte Carlo (MCMC) analyses can be monitored with the *coda* package.

Example 5. Open database connection. *RODBC* implements ODBC with compliant databases when drivers exist on the host system. The following is an example for reading all columns of tblOutput in an Microsoft® Access database *aedata.mdb*. The end result is a data frame called *tblOutput*.

```
# load the library and connect to Microsoft Access
library(RODBC)
c2 <- odbcConnectAccess("c:/aesop/mdb/aedata.mdb")
# select one table from the database
tblOutput <- sqlQuery(c2,paste("select * from tblOutput"))
# a data.frame
class(tblOutput)
```

This shows that R is able to make queries using structured query language (SQL) to a formal database system, so that marker information from genome-wide linkage and association studies can be organised and retrieved in a similar fashion and synchronised updates and communications are possible.

Comparison and integration with other software systems

As R has many functions available in a single environment, minimum effort is needed to write programs for data handling.

One can, then, concentrate on the statistical algorithm and analysis. This is clearly advantageous over stand-alone programs. The need to integrate internet capability within the data analytical system is also essential, given that data on several international projects are available from the internet.

Some researchers prefer to analyse data from large genetic studies using a hybrid of Perl or other scripts with programs written in C/C++; however, these programs are more targeted at computing professionals, having a relatively smaller statistical component. It may be more difficult to re-use codes written for such purposes. Program development may be more time-consuming, especially when analysis involving both genetic and environmental factors is required.[51] Even so, it is possible to use R as an independent program for such purposes. Likewise, compiled R packages for specific computer systems, but not the source code, can be distributed if necessary.

A large number of programs with a GUI have been developed recently. A notable example in multilevel modelling is the MIXOR/MIXREG and associated programs for longitudinal data analysis. With few exceptions, such as UNPHASE and JPAP, the source codes are largely unavailable, so it is sometimes difficult to assess the validity of the programs. Users should therefore remain familiar with a variety of implementations. They will encounter the usual problems of idiosyncratic data formats and source codes that are difficult to reuse. An alternative would be to use Java as an interface to the standard-alone programs, in order to run them in batch mode. The documentation associated with individual functions, however, is often poorer than those in R. In this regard, a different interface is provided by Rweb.[52] It provides a simple text entry form that returns output and graphs and a more sophisticated Javascript version that provides a multiple window environment and a set of point and click modules that are useful for introductory statistics courses and require no knowledge of the R language. All of the Rweb versions can analyse internet-accessible datasets if a URL is provided. It has also been shown that Perl can be used within R (see the *gregmisc* package).

Software development could be based on other environments — for example, Stata, SAS and S-PLUS, including some corporate efforts, such as SAS/GENETICS. The R package *foreign* provides commands to read and write dBase, Stata, SPSS and SAS xport files or access to Microsoft Excel/Access via ODBC, whereas data transformations between Stata and other applications require STAT/Transfer. Most programs written in R can be used with little alteration under S-PLUS. R has a clear advantage on graphics, and it is easier to incorporate routines written in C/C++/Fortran. Unlike SAS, it does not require a separate module for matrix operations.

A final note is given here regarding feedback that the authors received when developing *R/gap* and *kinship*, so as to show the benefit of the collaborative work that R encourages. The *hwe.hardy* function in *R/gap* was originally designed to accept only the full array containing the genotype counts, but was later extended according to a recommendation to use the *genotype* objects created by the *genetics* package. The C output format, *%lf*, was not supported by the American National Standards Institute (ANSI) standard and was subsequently changed following the advice of the R core development team. A compiling error with *emx.f* in the original POINTER program was also pointed out and later fixed. The *kinship* package was ported directly from S-PLUS. Extensive efforts were required for debugging; however, this has been greatly facilitated by the package *debug* from CRAN. There was also a problem with MacOS X in *kinship*, but this was subsequently changed according to suggestions.

Discussion

The authors have described both the motivation and prospects for using R as an integrated environment for genetic data analysis. While a formal presentation of R and comparison with R systems might have been given, the description has been deliberately kept informal. The following recaps the features of the R system.

First, it provides a flexible, integrated environment for statistical computing using an object-orientated programming language. It provides standard formats for data input, documentation and an interface to general statistical packages such as Stata, SPSS, SAS, S-PLUS and databases such as dBase, Microsoft Access/Excel, Oracle (http://www.oracle.com) and MySQL (http:/www.mysql.com). Above all, the R system is now a collaborative work, involving many people, and is available on most computer systems. Secondly, the environment can be greatly enhanced by contributed packages, which can either be implemented in the native R language or as a hybrid with external languages such as C/C++/Fortran/Perl. This allows for the easy incorporation of rich collections of algorithms and programs that have already been developed over the years. Packages can also be usefully incorporated from other areas of research. For example, packages for operations research, statistics in psychology, social network analysis, neuroimaging and spatial disease mapping are available in the same repository. Thirdly, standard datasets or benchmarks can be included as native objects in a package; these are ideal for evaluating new analytical methods. Fourthly, the functions and data in a package can serve for a variety of analyses. In haplotype analysis, for example, this could include estimation of haplotype frequencies, assignment of possible haplotypes, A linkage disequilibrium heat map and conditional and joint analysis with environmental factors, among others.

Given that the development of the R system is relatively recent, the wide range of tools available is impressive. The comprehensive and powerful features of R in data management, graphics and standard statistical analysis are making it a very useful platform for microarray data pre-processing, visualisation and advanced statistical analysis. There is also a rich array of packages for the analysis of population data and analysis, phylogenetic analysis and the analysis of quantitative traits from experimental design, although there is still a relative shortage of packages for the calculation of identify-by-descent and therefore of discrete traits or QTLs in human pedigrees. Packages for complex segregation analysis and path analysis are still experimental. Given the ease of creating packages from code that is already available, however, we expect that this situation will soon change.

Two important points should be made here. First, it should be pointed out that the use of R should not block the development of stand-alone programs. Secondly, a distinction should be made between potential and reality. The authors have come across arguments that implementations are trivial and that computer programming including R programming by statisticians are by default, straightforward. This may not be the case, however, and a more thoughtful approach is necessary. Often, software is cursorily written and poorly documented with no consideration for generality and use of examples, and consequently is hardly of any practical value. Fortunately, with the help of the R core development team, it is possible to produce industry-standard applications. The authors note that, at the time of writing, a special issue of the *Journal of Statistical Software* (http://www.jstatsoft.org) has been devoted to the transition of packages in XLISP-STAT to R. Given the current situation in genetic data analysis, it is now time for action.

In summary, the authors believe that R can potentially serve as an integrated platform for analysis of genetic data. While the packages currently available are limited in R, it is expected that its rich features will increasingly attract more developers and users. Further attention by theoretical and applied geneticists for software development and analysis will be very rewarding in the long term.

Acknowledgments

This paper was partly prompted by presentations to colleagues at King's College London and University College London. We are extremely grateful to Dr Mike Weale for his careful reading of the manuscript and for making many suggestions. The development of R packages was partly supported by the US National Institute of Aging (NIA) grant AG13196. Q.T. wishes to acknowledge support from NIA grant NIA-P01-AG08761 and 'The Micro-Array Center' project under the Biotechnological Research Program, financed by the Danish Research Agency. The authors wish to thank colleagues who generously make their software programs publicly available and provide useful feedback, whose contributions are well documented in *R/gap*. J.H.Z. wishes to thank John Kimmel from Springer for his encouragement in his work with R.

References

1. Collins, F.S., Morgan, M. and Patrinos, A. (2003), 'The Human Genome Project: Lessons from large-scale biology', *Science* Vol. 300, pp. 286–290.
2. Weale, M.E. (2004), 'A survey of current software for haplotype phase inference', *Hum. Genomics* Vol. 1, pp. 141–144.
3. Salem, R.M., Wessel, J. and Schork, N.J. (2005), 'A comprehensive literature review of haplotyping software and methods for use with unrelated individuals', *Hum. Genomics* Vol. 2, pp. 39–66.
4. Lander, E.S. and Schork, N.J. (1994), 'Genetic dissection of complex traits', *Science* Vol. 265, pp. 2037–2048.
5. Dudbridge, F. (2003), 'A survey of current software for linkage analysis', *Hum. Genomics* Vol. 1, pp. 63–65.
6. Ott, J. (1976), 'A computer program for linkage analysis of general human pedigrees', *Am. J. Hum. Genet.* Vol. 28, pp. 528–529.
7. Ott, J. (1974), 'Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies', *Am. J. Hum. Genet.* Vol. 26, pp. 588–597.
8. Morton, N.E., Rao, D.C. and Lalouel, J.-M. (1983), 'Methods in Genetic Epidemiology', Karger, New York, NY.
9. Guo, S.W. and Lange, K. (2000), 'Genetic mapping of complex traits: Promises, problems, and prospects', *Theor. Popul. Biol.* Vol. 57, pp. 1–11.
10. Venables, W.N. and Ripley, B.D. (2002), 'Modern Applied Statistics with S', Springer, New York, NY.
11. Chambers, J.M. (1998), 'Programming with Data', Springer, New York.
12. Chambers, J.M. and Hastie, T.J. (1992), 'Statistical Models in S', Chapman & Hall, London, UK.
13. Satagopan, J.M., Yandell, B.S., Newton, M.A. *et al.* (1996), 'A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo', *Genetics* Vol. 144, pp. 805–816.
14. Burkett, K., McNeney, B. and Graham, J. (2004), 'A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models', *Hum. Hered.* Vol. 57, pp. 200–206.
15. Schaid, D.J., Rowland, C.M., Tines, D.E. *et al.* (2002), 'Score tests for association between traits and haplotypes when linkage phase is ambiguous', *Am. J. Hum. Genet.* Vol. 70, pp. 425–434.
16. Lake, S.L., Lyon, H., Tantisira, K. *et al.* (2003), 'Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous', *Hum. Hered.* Vol. 55, pp. 56–65.
17. Goudet, J. (2005), 'Hierfstat, a package for R to compute and test variance components and F statistics', *Mol. Ecol. Notes* Vol. 5, pp. 184–186.
18. Weir, B.S. and Wilson, S.R. (1986), 'Log-linear models for linked loci', *Biometrics* Vol. 42, pp. 665–670.
19. Huttley, G.A. and Wilson, S.R. (2000), 'Testing for concordant equilibria between population samples', *Genetics* Vol. 156, pp. 2127–2135.
20. Luo, Z.W. (1998), 'Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations', *Heredity* Vol. 80, pp. 198–208.
21. Sen, S., Satagopan, J. and Churchill, J. (2005), 'QTL study design from an information perspective', *Genetics* Vol. 170, pp. 447–464.
22. Devlin, B. and Roeder, K. (1999), 'Genomic control for association studies', *Biometrics* Vol. 55, pp. 997–1004.
23. Hirotsu, C., Aoki, S., Inada, T. *et al.* (2001), 'An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis', *Biometrics* Vol. 57, pp. 769–778.
24. Zhao, J.H. (2004), '2LD, GENECOUNTING and HAP: Computer programs for linkage disequilibrium analysis', *Bioinformatics* Vol. 20, pp. 1325–1326.
25. Yu, C. and Zelterman, D. (2002), 'Statistical inference for familial disease clusters', *Biometrics* Vol. 58, pp. 481–491.
26. Strand, A. (2002), 'Metasim 1.0: An individual-based environment for simulating population genetics of complex population dynamics', *Mol. Ecol. Notes* Vol. 2, p. 376.
27. Broman, K.W., Wu, H., Sen, S. *et al.* (2003), 'R/qtl: QTL mapping in experimental crosses', *Bioinformatics* Vol. 19, pp. 889–890.

28. Mott, R., Talbot, C.J., Turri, M.G. *et al.* (2000), 'A method for fine mapping quantitative trait loci in outbred animal stocks', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 12649–12654.
29. Clayton, D. and Jones, H. (1999), 'Transmission/disequilibrium tests for extended marker haplotypes', *Am. J. Hum. Genet.* Vol. 65, pp. 1161–1169.
30. Nicholson, G., Smith, A.V., Jónsson, F. *et al.* (2002), 'Assessing population differentiation and isolation from single-nucleotide polymorphism data', *J.R. Stat. Soc. B* Vol. 64, pp. 695–715.
31. Espeland, M.A. (1986), 'A general class of models for discrete multivariate data', *Commun. Stat.-Simul.* Vol. 15, pp. 405–424.
32. Friendly, M. (2000), 'Visualizing Categorical Data', SAS Institute, Cary, NC.
33. Efron, B. (2004), 'Large-scale simultaneous hypothesis testing', *J. Am. Stat. Assoc.* Vol. 99, pp. 96–104.
34. Tritchlern, D., Liu, Y. and Fallah, S. (2003), 'A test of linkage for complex discrete and continuous traits in nuclear families', *Biometrics* Vol. 59, pp. 382–392.
35. Liu, Y., Tritchler, D. and Bull, S.B. (2002), 'A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits', *Genet. Epidemiol.* Vol. 22, pp. 26–40.
36. Yang, Y., Zhang, J., Hoh, J. *et al.* (2003), 'Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA', *Proc. Natl. Acad. Sci. USA* Vol. 100, pp. 7225–7230.
37. Betensky, R.A. and Rabinowitz, D. (2000), 'Simple approximations for the maximal transmission/disequilibrium test with a multi-allelic marker', *Ann. Hum. Genet.* Vol. 64, pp. 567–574.
38. Chapman, J.M., Cooper, J.D., Todd, J.A. *et al.* (2003), 'Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power', *Hum. Hered.* Vol. 56, pp. 18–31.
39. Lowe, C.E., Cooper, J.D., Chapman, J.M. *et al.* (2004), 'Cost-effective analysis of candidate genes using htSNPs: A staged approach', *Genes Immun.* Vol. 5, pp. 301–305.
40. Zapata, C., Carollo, C. and Rodriguez, S. (2001), 'Sampling variance and distribution of the D measure of overall gametic disequilibrium between multiallelic loci', *Ann. Hum. Genet.* Vol. 65, pp. 395–406.
41. Zhao, J.H., Curtis, D. and Sham, P.C. (2000), 'Model-free analysis and permutation tests for allelic associations', *Hum. Hered.* Vol. 50, pp. 133–139.
42. Zhao, J.H. and Sham, P.C. (2002), 'Faster haplotype frequency estimation using unrelated subjects', *Hum. Hered.* Vol. 53, pp. 36–41.
43. Zhao, J.H., Lissarrague, S., Essioux, L. *et al.* (2002), 'GENECOUNTING: Haplotype analysis with missing genotypes', *Bioinformatics* Vol. 18, pp. 1694–1695.
44. Excoffier, L. and Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.* Vol. 12, pp. 921–927.
45. Noonan, J.P., Li, J., Nguyen, L. *et al.* (2003), 'Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin alpha cluster', *Am. J. Hum. Genet.* Vol. 72, pp. 621–635.
46. Zaykin, D.V., Westfall, P.H., Young, S.S. *et al.* (2002), 'Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals', *Hum. Hered.* Vol. 53, pp. 79–91.
47. Kerem, B.S., Rommens, J.M., Buchanan, J.A. *et al.* (1989), 'Identification of the cystic fibrosis gene: Genetic analysis', *Science* Vol. 245, pp. 1073–1080.
48. Daly, M.J., Rioux, J.D., Schaffner, S.E. *et al.* (2001), 'High-resolution haplotype structure in the human genome', *Nat. Genet.* Vol. 29, pp. 229–232.
49. Curtis, D., Zhao, J.H. and Sham, P.C. (1999), 'Comparison of GENEHUNTER and MFLINK for analysis of COGA linkage data', *Genet. Epidemiol.* Vol. 17(Suppl. 1), pp. S115–S120.
50. Zhao, J.H. (2005), 'Mixed-effects Cox models of alcohol dependence in extended families', *BMC Genet.*, In press.
51. Merikangas, K.R. and Risch, N. (2003), 'Will the genomics revolution revolutionize psychiatry?', *Am. J. Psychiatry* Vol. 160, pp. 625–635.
52. Banfield, J. (1999), 'Rweb: Web-based statistical analysis', *J. Stat. Soft.*, Vol. 4, pp. 1–15.