Genetic Dissection of Complex Traits *In Silico*: Approaches, Problems and Solutions

Jing Hua Zhao[*,1] and Qihua Tan[2,3]

[1]*MRC Epidemiology Unit, UK*

[2]*Odense University Hospital, Denmark*

[3]*Institute of Public Health, University of Southern Denmark*

Abstract: The genome projects in human and other species have made genetic data widely available and pose challenges as well as opportunities for statistical analysis. In this paper we elaborate the concept of integrated analysis of genetic data, such that most aspects of analyses can be done effectively and efficiently in environments with facility for database accessibility, graphics, mathematical/statistical routines, flexible programming language, re-use of available codes, Internet connectivity and availability. This extends an earlier discussion on software consolidation (Guo and Lange. Theor Pop Biol 57:1-11, 2000). A general context is laid out by recollecting the research paradigms for genetic mapping of complex traits and illustrated with the study of ageing, before turning to the computational tools currently used. We show that the R system (http://www.r-project.org) so far is the most comprehensive and widely available system. However, other commercial systems can potentially be successful. In particular, we compare SAS (http://www.sas.com), Stata (http://www.stata.com), S-PLUS (http://www.insightful.com) and give some indications of future development. Our investigation has important implications for both statisticians end other researchers actively engaged in analysis of genetic data.

Keywords: Complex traits, genetic epidemiology, integrated analysis, statistical and computational genetics.

1. INTRODUCTION

With successful mapping and localisation of susceptible genes for Mendelian diseases such as Huntington's disease [1], cystic fibrosis [2] and some form of non-Mendelian diseases such as breast cancer [3], much of the current research in human genetics is focused on common diseases including asthma, cardiovascular diseases, diabetes, and psychiatric disorders, among others. The difficulty with these "complex traits" has provoked the close scrutiny of various aspects of research [4, 5] which involves study design, statistical analysis, and gene-environment interaction. To a large extent, methods underpinning these efforts have been the subject of active research by mathematicians and statisticians over a century. However, never than before are they under increasing challenges, predominantly due to the fact that the genetic mechanisms for common diseases in relation to environment are much more complicated and not well-understood. This understanding calls for better knowledge of human demographic history, biological pathways and physiological functions, which relies on whole genomics data as well as large collections of biological and environmental covariates. The international HapMap project [6, 7] is one example for study of patterns of common DNA variations [8-10] across the genome using millions of single nucleotide polymorphisms (SNPs) [11], which are the most abundant form of genetic variation and accounts for about 90% of human DNA polymorphism. Studies of SNPs over large genomic regions therefore hold great promise for mapping polygenic disease loci, as has been demonstrated [12]. They can further shed light on human history including relationships between ethnic groups, migrations, together with evolutionary information such as genetic drift, selection, mutation, and recombination at the molecular level. The explanation of the frequency and distribution of these SNPs throughout the genome has thus been regarded as a central challenge [12, 13]. The recognition of the need for population studies results in effort on a national level such as UK Biobank initiative (http://www.ukbiobank.ac.uk) and similar projects envisaged elsewhere [14]. These projects represent initiatives to collect a large array of genetic and environmental factors and will likely be longitudinal. Apart from human genome project, many "*-omics*" projects are carried out and paralleled in other species. Additional account of the genetic databases is available on the first issue of Nucleic Acids Research each year. Greater complexity implicit in the data also calls for more sophisticated modelling. Data mining and modelling of socio-biological pathways while accounting for socio-historical events are likely to be ubiquitous. Therefore further data fusion is expected.

The annotation of the large genomic data currently is an unwieldy task. For instance, genome-wide association studies involving several thousand individuals each with tens of thousands of SNPs have been planned or carried out for most common diseases [15]. This would thwart most of the traditional analytical tools currently in use. Not only is greater computing power needed, but classic theories such as those for multiple testing need to be re-examined. While many researchers continue struggling with an increasing and bewildering number of computer programs for their study design, analysis and reporting, as shown in several recent reviews [16-20], there is an urgent need for concerted effort

*Address correspondence to this author at the MRC Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK; Tel: +44 1223 741398; Fax: +44 1223 740050;
E-mail: jinghua.zhao@mrc-epid.cam.ac.uk

1574-8936/06 $50.00+.00

to produce software systems for analysis of genetic data. Earlier, it was indicated that *"Many statistical geneticists sense that the time is ripe for software consolidation. A critical mass of users to support commercialization is now in place, particularly with the entry of the pharmaceutical industry into genetic epidemiology. The problem is that no one is sure how to achieve software consolidation. Most statisticians have their pet methods, which they are loath to give up. Nonetheless, there are some clear demands that will drive the process of software development. Among the obvious needs are (1) a professional-looking graphical user interface, (2) integration of genetic databases and analysis programs, (3) integration of visual display of pedigree data with analysis output, (4) better analysis tools for screening pedigree data for genotyping errors, (5) automatic choice of the quickest algorithms for likelihood evaluation, and (6) flexible programs that allow sophisticated users to pursue novel models and variations on existing models"* [21].

We feel two points are outstanding, which are a clear correspondence between research paradigm of genetic investigation and the analytical or computing machinery, and the need from end users for software consolidation. There have been a lot of discussions on the statistical methodology (e.g., [22-30]) but relatively few on computing tools. An attempt was made on integrated analysis of genetic data using the R system (http://www.r-project.org) [31], and we will elaborate related software systems and issues discussed there. We remain our focus on the problems and prospect of data analytic practice on human genetics, particularly on genetic epidemiology. In the discussion we will give some indication as to how integrated systems for genetic data could be developed.

2. APPROACHES TO GENETICS OF COMPLEX TRAITS

No universal definition of complex trait exists, which broadly refers to any phenotype that does not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene locus [22], Examples are susceptibilities to asthma, cardiovascular diseases, diabetes, cancer, infection and psychiatric disorders. The lack of simple genotype-phenotype relationship lies not only in complexities of physiology and population but also the environment. These complexities can lead to a number of specific phenomena including polygenic inheritance, locus heterogeneity, epistasis or gene-gene interaction, environmental vulnerability, gene-environment interactions, development or time-dependent expression of genes, general aging of the system [4, 32]. Other characteristics of complex trait have been documented, e.g., birth order and cohort effects, late or variable age of onset, and variable disease progression. Many complex diseases are hard to diagnose accurately or with measurement errors for quantitative traits [33]. There is also limited statistical power to detect genes that are involved [24, 34]. The study of the genetic predisposition underlying these diseases has been the central theme for genetic epidemiology [35-41], *"the study of the joint action of genes and environmental factors in causing disease in human population and their patterns of inheritance in families"* [28] or similarly *"a science that deals with the aetiology, distribution and control of disease in groups of relatives and with inherited causes of disease in population... genetic epidemiology should be broadly defined to embrace all aspects of population genetics except evolution, including phenotypes, gene-environmental interactions, and modes of transmission related to health or to location of disease genes, or requiring methods of analysis developed for genetic determinants of disease... other disciplines are subsumed by this definition, including ecogenetics, behaviour genetics, and demographic genetics ... molecular epidemiology as it relates to inherited risk factors is contained within genetic epidemiology"* [39]. This is in contrast with general epidemiology studying the distribution, determinants [and control] of health-related states and events in population [41]. A recent account of the analytic strategy is given in [42]. Traditional strategies to assess the genetic influences of diseases include twin and adoption studies to characterise the relative contributions of genetic and environmental components, the study of familial aggregation of diseases (e.g. path analysis [43]), the modality and mode of inheritance of major gene (e.g. the commingling [44] and segregation analyses [45]), linkage [46-49] and association analysis [50-52]. Before the 1980s, there were only a limited number of genetic markers available therefore restricting the scope of linkage and association studies, only to be changed by the human genome project. Currently, the mapping methods have been classified into several categories, namely parametric LOD score methods based on assumed disease models, variance component methods for quantitative traits, allele-sharing methods, association studies, construction of conserved haplotypes and mapping *via* experimental species [33]. The process of genetic epidemiology has recently reframed [41, 53-58] as with the following steps: descriptive epidemiology, familial epidemiology, segregation analysis, linkage analysis, fine mapping, association with candidate genes, cloning the gene and identifying mutations, characterising the gene [28]. Note this encapsulates the commonly used candidate genes and genome-screen approaches [4, 32, 59-61], particularly the so-called genome-wide association studies [15, 62-64].

The genetic epidemiology of human ageing and longevity serves as an illustrative example and is sketched elsewhere [65]. The lifespan familial correlation was described in 1899 [66] and continued to be examined from then on. Data from well-defined populations support the notion that there exist transmittable familial attributes, which are partly genetic, affecting lifespan. A recent report suggested an exceptional genetic inheritance due to a high concentration of long-lived ancestors [67]. Twin studies were used to estimate the genetic and environmental components to human survival, and a large Danish twin study estimated that the genetic component accounts for about 25% of the total lifespan variation [68]. A further step following the confirmation of genetic influence in lifespan is to look for the causal genetic variants. The past few decades have witnessed extensive use of genetic data [69]. The first genome-wide linkage analysis on human exceptional longevity used sample of 137 families with exceptional longevity [70] to identify a region on chromosome 4 that could possibly harbor a gene affecting human longevity. Among association studies using candidate genes, the apolipoprotein E gene has given the most reproducible and strongest association and its emerging mechanisms in the etiology of heart disease, stroke and Alzheimer's disease which are among the major threats to

public health [71]. Furthermore, gene-expression studies have also been carried out [65].

3. THE PRACTICAL DIFFICULTY OF GENETIC DATA ANALYSIS

The practice of genetic epidemiology is greatly influenced by advance in modern computing and molecular biology. The range of analyses carried out by genetic epidemiologist has largely been through successful uses of many software programs. In the early 1970s, this was represented by algorithmic breakthrough in pedigree analysis [72] and computer implementation [47]. In the early 1980s most software programs were largely outlined in [73]. This was followed by development in algorithm and software was seen in the late 1980s [48, 49, 74], 1990s [75-82] and more recently [83-85]. A current list of software programs for genetic epidemiology is maintained at http://www.nslij-genetics.org/soft/ and mirrored at http://linkage.rockefeller.edu/soft/. References to most of the packages given below can be found there and therefore omitted here. A somewhat reduced set of programs from the former UK HGMP-RC (Human Genome Mapping Project Resource Centre) is freely available on LITBIO (http://www.litbio.org). Surveys have been given of software on linkage analysis [16], haplotype phase inference [18], on tag SNPs [86], and genetic power calculation [17]. A survey of programs for haplotype analysis [19] contained 46 software programs. A comparison of linkage analysis for quantitative trait [20] includes LINKAGE, FASTLINK, PAP, SOLAR, SEGPATH, ACT, Mx, MERLIN, GENEHUNTER, Loki, Mendel, SAGE, QTDT and FBAT. These programs implement variance components, Markov chain Monte Carlo (MCMC), Haseman–Elston [87], penetrance model-based linkage analyses, as well as measured genotype association analyses and quantitative trait transmission disequilibrium tests (TDTs).

While proved useful in specific applications, they are also scattered and often hastily written. They were written in many computer languages, ranging from C, C++, Fortran, Pascal, Java, Perl to standard statistical packages; some of which only available in compiled form and tested under specific computer systems. They require data in specific formats, often not conforming to any standard, and cumbersome to reuse output from these programs. Some include simple parsing and few have graphical facility. In analysis of large data, it is often necessary to write customised utilities for these programs. The range of skills required for different computer languages and tools largely requires a computer professional. They often lead to redundant work, poor maintenance, and lack of validity checks. Further disadvantage of these packages is that they were designed from the developers own experiences, therefore require a variety of data format and difficult to adapt by other users. They are often platform-specific, and require other programs for accessing database, for elementary statistical computing and inference, for graphics, and for Internet connectivity.

Attempts have been made to resolve these difficulties by comprehensive implementation, e.g. SAGE (Statistical Analysis for Genetic Epidemiology) [85], Mendel [74], MORGAN (Monte Carlo Genetic Analysis), PAP (Programs for Pedigree analysis) [88] are all quite comprehensive. Packages for population genetic analyses of molecular data [89] include GENEPOP (http://wbiomed.curtin.edu.au/genepop/), GDA (Genetic Data Analysis) [90], Arlequin [91] and POPGENE (http://www.ualberta.ca/~fyeh/). However, most have their limitations, and many aforementioned problems remain. It is even more difficult to adapt these programs.

4. THE EMERGING ALTERNATIVES

Recently, components of genetic data begin to appear on commercial packages to various degrees, including SAS (http://www.sas.com), Stata (http://www.stata.com), S-PLUS (http://www.insightful.com) and Genstat (http://www.vsn-intl.com/genstat/). A non-commercial alternative is the GNU (http://www.gnu.org) version of S-PLUS called R [92] (http://www.r-project.org), available on Unix/Linux, Windows, MacOS X and a number of other computer systems. The R system can be considered as part of the open source initiative (http://www.opensource.org), which includes operating system (e.g. Linux), software (e.g. Apache server, Fortran/C/C++ compilers), database management system (e.g. MySQL, http://www.mysql.com, and Oracle, http://www.oracle.com) and scripting languages (e.g. Perl, http://www.perl.org), all freely available. These systems all have comprehensive procedures for statistical data analysis, professional graphic user interfaces and ability to access database directly or through open database connectivity (ODBC). It is possible to call commands in the host computer system *via* **system** or **shell** from these packages, making it possible to run stand-alone programs and retrieve their output. It is possible to call programs in C/C++/Fortran, through SAS/TOOKIT in SAS, .Call/.Internal/.C/.Fortran in R but not possible with Stata. We give a brief summary of these systems below, with the main features of these software systems highlighted in Table **1**. As it would be unrealistic to illustrate many features of these packages, we only illustrate their use of large dataset in the wake of whole-genome association studies. The example dataset is downloaded from the HapMap website in ASCII format, converted to Excel and Access, and uploaded to a MySQL server through ODBC. It is stored as a table named Genotypes_chr1 1_CEU in a MySQL database called test.

4.1. SAS

It has been one of the most powerful statistical systems available and consists of many modules. Briefly, SAS/BASE offers many procedures for data management including PROC SQL, while SAS/STAT provides procedure for most statistical analysis and SAS/GRAPH for computer graphics. Furthermore, SAS/IML provides an interactive matrix language, mathematical and statistical procedures which would be appropriate for statistical modelling. Other modules include SAS/ASSIST, SAS/ETS, SAS/GIS, SAS/OR, SAS/QC for menu-driven directive, econometric time-series analysis and forecasting, geographic information system, operations research, and industrial process control. Finally, SAS/ACCESS, SAS/SHARE, SAS/CONNECT have facility for database access and connection between software systems. While over years many macros have been written for genetic analysis, a recent synthesis is available through its new SAS/GENETICS module, which contains

the following procedures for genetic association analysis: ALLELE, HAPLOTYPE, CASECONTROL, HTSNP, FAMILY, INBREED, PSMOOTH. These procedures cover analysis ranging from summary statistics to analysis of population and family data. Procedures in SAS/STAT include MULTTEST for multiple testing and GLMM for generalised linear mixed modelling.

We now use an example from a collaborative study of the linkage disequilibrium in eleven population isolates, a dataset of 100 family trios on chromosome 22. We wish to examine if there is evidence of transmission distortion on a sample from Antioquia, Colombia. There were 2696 SNPs genotyped in the dataset, organised by individual ID, SNP number and their alleles. This could be done with TDT implemented in the PROC FAMILY. The test is furnished with the following program.

The program starts by reading an ASCII-formatted file ant.txt containing marker names and alleles, and preparing the IDs for later analysis. This is followed by descriptive analysis using PROC ALLELE for marker and genotype frequencies, Hardy-Weinberg equilibrium tests. As there are many markers involved, we resort to the output delivery system (ODS) facility to keep marker information including allele and genotype frequencies as datasets for other purposes. Later, PROC FAMILY indicates Mendelian inconsistencies of genotypes in the data. The p values were obtained according to the so-called reconstruction TDT [93] and passed to PROC MULTTEST to calculate the Hochberg's [94] and Benjamini and Hochberg's step-up methods for p-value adjustments [95]. Other options such as SDT [96] and STDT [97] are possible and can be combined with COMBINE option. The permutation tests are also possible with PERMS option. The results conform to a

```
/*to set up trio information*/
data ant;
     infile 'ant.txt' firstobs=2 dlm=",";
     attrib marker length = $12.;
     input eid$ ant$ marker$ a1$ a2$;
     pid=substr(eid,1,6);
     id=substr(eid,8,1)+0;
     if id=3 then do; disease=2; fid=1; mid=2; sex=1; end;
     else do; disease=1; fid=0; mid=0;
            if id=1 then sex=1; else sex=2;
          end;
     drop ant;
run;
/*to sort data by marker order*/
proc sort data=ant;
     by marker;
run;
/*to perform marker-marker analysis*/
ods select none;
proc allele data=ant;
     ods output markersumm=ms
                allelefreq=af genotypefreq=gf;
     where id ^=3;
     by marker;
     var a1 a2;
run;
ods select all;
proc print data=ms;
title Marker summary information;
proc print data=af;
title Allele frequencies;
proc print data=gf;
title Genotype frequencies;
run;
/*to perform TDTs*/
proc family data=ant prefix=Marker rctdt outstat=p;
     by marker;
     id id fid mid;
     var a1 a2;
     trait disease / affected=2;
run;
proc print data=p;
run;
/*distribution of p-values*/
proc univariate data=p;
     histogram probRCTDT;
run;
data p;
     set p;
     test=compress('test'|| n_);
     rename probRCTDT=raw_p;
run;
/*to obtain adjusted p-values*/
proc multtest pdata=p fdr hoc;
run;
```

similar analysis through UNPHASED [98], but with much simpler programming effort. Note for most standalone programs for TDT, it is necessary to create a standard dataset with records for all members in a trio even when some members are unavailable due to lack of genotype data. The whole analysis would have been quite clumsy for basic marker information and TDTs.

The following program shows how ODBC and MySQL engines could be used to access table Genotypes_chr11_CEU.

```
libname test odbc datasrc=myodbc user=jhz22;
proc print data=test.Genotypes_chr11_CEU;
run;
proc sql;
     connect to odbc as test (datasrc=myodbc user=jhz22);
     select * from test.pet;
libname test2 mysql database=test user=jhz22;
proc print data=test2.pet;
run;
proc sql;
     connect to mysql as test2 (database=test user=jhz22);
     select * from test2.pet;
```

The MicroSoft Access database can be accessed directly under Windows using IMPORT procedure,

```
%let dbname = c:\hapmap\db1.mdb;
%let uid = xxxx;
%let pwd = *****;
%let wgdb = c:\hapmap\db1.mdw;
proc import out = objects
     datatable = "Genotypes_chr11_CEU"
     dbms = access97 replace;
     database = "&dbname";
     userid = "&uid";
     password = "&pwd";
     wgdb = "&wgdb";
run;
```

4.2. Stata

It is fast becoming the most popular computer package for epidemiologists, and favoured by researcher in other fields such as econometrics [99]. It has facilities ranging from basic data management, computer graphics, programming, to methods for complex survey, longitudinal data analysis and multilevel models. Beside its simple but flexible syntax and comprehensive on-line documentation, it has many unique features such as frequency, probability and analytic weights. Not surprisingly, there are efforts to implement genetic analysis (e.g. http://www.cimr.cam.ac.uk and biostatistics resources, http://www.biostat-resrouces.com), though so far no corporate initiative is involved. To show the ease to install Stata packages from the Internet, the popular package by David Clayton for haplotype tagging [100] can be downloaded as follows,

```
. net from http://www-gene.cimr.cam.ac.uk/clayton/software/stata
. net describe htSNP2
. net install htSNP2
. describe using http://www.stata-press.com/data/r8/clogitid.dta
. use http://www.stata-press.com/data/r8/clogitid.dta
```

The first three commands obtain a description of packages and install htSNP2. The remaining two commands access the example data set for conditional logistic regression from Stata's homepage. Amendments to any command can be achieved with the Stata command **adoedit**.

Several other packages are also available from Biostatistics resources, e.g.,

```
. net from http://www.biostat-resources.com/stata
```

for the list of packages including genhw, gencc, qtlsnp, hwsnp, genecmt for Hardy-Weinberg equilibrium test, case-control and case-parent-triad data analysis, etc. It is self-evident how these packages are distributed from the authors' websites.

The following is the Stata code to query the MySQL database:

```
. odbc list
Data Source Name                         Driver
------------------------------------------------------------
myodbc                                   MySQL ODBC
3.51 Driver
------------------------------------------------------------
. odbc query "myodbc"
DataSource: MySQL database
Path      : jhz22
------------------------------------------------------------
Genotypes_chr11_CEU
------------------------------------------------------------
. odbc desc "Genotypes_chr4_HCB",
dialog(complete)
. set mem 50M
. odbc load, exec("select * from
Genotypes_chr11_CEU")
```

The common approach for marker-trait association, the so-called haplotype trend regression [121], is particularly simple with the pweight using constructed haplotypes from any haplotype construction programs, e.g. **logit cc locus* [pweight=probability]** specifies a logistic regression model of a binary outcome (cc) with haplotypes (locus*) weighted by the posterior probabilities (probability).

Table 1. A summary of Main Features in the Four Environments (SAS, Stata, S-PLUS and R)

Name	Comprehensive Analytic procedures	Interactive and batch mode	Graphical facility	Database connectivity	Extensibility	Documentation	Ability to C/Fortran routines	Internet function	Number of packages for genetic analysis	Cost
SAS	Yes	Yes	Yes	Yes	Yes	Complete	SAS/Toolkit or shell	Yes	Mostly in SAS/GENETICS	Expensive and annual licence
Stata	Yes	Yes	Yes	Yes	Yes	Partly online	Shell	Yes	Relatively small	Relatively cheap
S-PLUS	Yes	Yes	Yes	Yes	Yes	Complete	Yes	Yes	Relatively small	Moderate
R	Yes	Yes	Yes	Yes	Yes	Complete	Yes	Yes	Large	Free

Note. The analytic procedures range from numerical analysis and probability distributions to linear and other models in modern applied statistics. In SAS and S-PLUS some procedures are available through additional module such as SAS/OR for operations research, S-PLUS NuOPT and S+ArrayAnalyzer for numerical optimisation and microarray analysis. In R, several packages are devoted to numerical optimisation and Markov chain Monte Carlo. SAS has established facility for graphic information system while S-PLUS achieves this through Arcview. Stata is the only one without making full manuals available electronically.

4.3. S-PLUS

It is based on the S language [101] and designed for data handling, analysis and graphics [102], with facility to handle large dataset in its latest version. Its flexibility in graphics lies in little effort to create and add items to plots with calls to lines, polygon, etc. and to use toolbox containing many functions. Collection of function and data is in the form of package to be called with **library**() command when necessary. It also has S+ArrayAnalyzer for analysing microarry data [103], Taqman analysis, and packages haplo.stats [104], kinship [105, 106], multic [107] and multigene [108].

Here is the S-PLUS code to access the MySQL database; the current S-PLUS version 7 on Linux does not support ODBC.

```
library("S-MySQL",lib.loc="/home/jhz22/S/library")
# initialize S-PLUS as a MySQL client
mgr <- dbManager("MySQL")
# create a connection to a MySQL server
con <- dbConnect(mgr, user="jhz22", dbname="test")
# run a query, leave results on the server
rs <- dbExec(con, "select * from Genotypes_chr11_CEU")
# fetch up to, say, 5 records
df <- fetch(rs, n = 5)
# close resultSet rs and connection con
close(rs)
close(con)
```

The S-MySQL library can be obtained from http://stat.bell-labs.com/RS-DBI/download/.

4.4. R

Like in S-PLUS, functions and datasets in R can be organised as objects in specific packages. The large repository of packages is maintained at CRAN (Comprehensive R Archive Network, http://cran.r-project.org). Packages serving on similar purpose can further be grouped into CRAN task view (ctv). The number of procedures in R is much larger than S-PLUS and most packages developed for S-PLUS, including haplo.stats, kinship, multigene, are already ported to R. Among many features, R offers interface to databases such as MySQL and Oracle, MS Access as well as spreadsheet such as MS Excel. The package foreign is recommended as it has capability to import data from dBase and Stata, or SAS when the SAS system is available. R allows for graphic user interface be developed using its TCL/Tk (http://www.tcl.tk) component ---- an example is the Rcmdr package which implements a menu-driven interface for many analyses in R. The R package snow (Simple Network of Workstations) implements a simple mechanism for using a workstation cluster in R. The interface, which is based in part on the Python CoW (Cluster of Workstations) package, is intended to be quite simple, and is designed so that it can be implemented on top of several different lower level communication mechanisms. Three low level interfaces have

been implemented, one based on sockets, one using PVM *via* the rpvm package by Li and Rossini, and one using MPI, *via* the Rmpi package by Hao Yu. An example of using the cluster for parallel bootstrapping is given by Luke Tierney (http://www.stat.uiowa.edu/~luke/R/cluster/cluster.html).

R packages are usually obtained using the **download.packages**() command but R programs are loadable directly from the Internet using the **source**() command, e.g. **source**("http://wpicr.wpic.pitt.edu/WPICCompGen/genomic_control/gc.txt") for genomic control [109] and **source** ("http://www.uib.no/smis/gjessing/genetics/software/haplin/HAPLIN.BETA.R.txt") for case-parent trio data [110, 111]. The following is adapted from the CRAN task view on several categories of genetic data analyses.

4.4.1. Population Genetics

Package genetics contains classes and methods for representing genotype and haplotype data, with functions for population genetic analysis such as estimation and testing of Hardy-Weinberg and linkage disequilibria. Geneland has functions for detecting spatial structures from genetic data within a Bayesian framework *via* MCMC estimation. Malmig implements Malecot migration model and related functions. rmetasim provides an interface to the metasim engine for population genetics simulations. hapsim simulates haplotype data with pre-specified allele frequencies and linkage disequilibrium (LD) patterns. A few population genetics functions are also implemented in gap. hierfstat allows the estimation of hierarchical F-statistics from haploid or diploid genetic data. LDheatmap creates a heat map plot of measures of pairwise LD using D' or r. hwde fits models for genotypic disequilibria. Biodem package provides functions for biodemographical analysis, e.g. **Fst**() for Fst from the conditional kinship matrix. Package kinship offers some functions **kinship**(), **lmekin**(), and **coxme**() for analysis of survival data on large and extended pedigrees.

4.4.2. Phylogenetics

These include packages ape and apTreeshape for handling of phylogenetic trees and evolution analysis. Package ouch provides Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses, while phyloarray offers functions for phylogenetic microarray data processing. PHYLOGR is a suite of functions for the analysis of phylogenetically simulated data sets and model fitting. stepwise implements a method for stepwise detection of recombination breakpoints in sequence alignments.

4.4.3. Linkage, LD and Haplotype Mapping

Package gap contains functions for sample size calculations, probability of familial disease aggregation, kinship calculation, and some tests for linkage and association analyses. Among the other functions, **genecounting**() estimates haplotype frequencies from genotype data with missing values and applicable for both autosomal and X chromosome data. For family data, tdthap offers an implementation of TDT for extended marker haplotypes, whereas powerpkg performs power analyses for the affected sib pair and the TDT design. The package hapassoc performs likelihood inference of trait associations with haplotypes in generalised linear models. Package haplo.stats implements association tests for a wide variety of traits (e.g. binary, ordinal, quantitative, and Poisson). Its functions **haplo.em**() provided maximum likelihood estimation of haplotype probabilities and **haplo.glm**() for modelling gene-environment interactions. All packages above work when haplotype phase is uncertain. ldDesign is a package for design of experiments for association studies for detection of linkage disequilibrium.

4.4.4. QTL Mapping

They contain methods for the analysis of experimental crosses to identify markers contributing to variation in quantitative traits. Package bim is for Bayesian interval mapping diagnostics. bqtl implement both likelihood-based and Bayesian methods for inbred crosses and recombinant inbred lines. qtl provides several functions and a data structure for QTL mapping, including a function scanone() for genome-wide scans. The package qtlDesign has functions for designing QTL experiments, including power computations.

4.4.5. Multiple Testing

The package qvalue implements false discovery rate *via* function **qvalue**() [112], to be called either as a function or from a graphic interface. Package multtest [113, 114] also offers several non-parametric bootstrap and permutation resampling-based multiple testing procedures. There are additional packages such as locfdr [115, 116], twilight [117].

Many packages have not gone through the formal check by or submitted to CRAN, e.g. dgc.genetics (http://www-gene.cimr.cam.ac.uk/clayton/software/), happy (http://www.well.ox.ac.uk/happy/), hapgen and popgen (http://www.stats.ox.ac.uk/~marchini/), migration (http://www.math.ntnu.no/~jarlet/migration/) [118] and multic [107]. Other packages such as Bradley-Terry, epitools, evd, gllm, rmeta, vcd are quite general but with functions for genetic data. Routines for annotation of biological pathways [119, 120] in R are also attractive. Further advantage of R is its collection of packages for Bayesian data analysis, such as MCMCpack, mcmc, coda, boa, and interface to WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml).

The following code shows loading MySQL directly and MicroSoft Access databases through ODBC:

```
# MySQL
library(RMySQL)
m <- dbDriver("MySQL")
con <- dbConnect(m,"test")
rs <- dbSendQuery(con,"select * from
Genotypes_chr11_CEU")
df <- fetch(rs,n=3)
# ODBC
library(RODBC)
c2 <- odbcConnectAccess("db1.mdb")
# select the table
tblOutput <- sqlQuery(c2,paste("select
* from Genotypes_chr11_CEU"))
# the property of tblOutput
class(tblOutput)
```

5. DISCUSSION

To echo our earlier discussion, the large genomic data pose immense challenges for inter-disciplinary research between biologists, epidemiologists, mathematicians, statisticians, computer scientists and even sociologists. In particular, just as modern statistics is greatly influenced by advances in computing technology, these challenges equally mean opportunities. We feel a sensible strategy would be to "integrate" genetic data analysis with "general-purpose" software systems currently available. We have explored several such systems for genetic data analysis. In accordance with the review by Guo and Lange [21], we note that these indeed have some degree of commercialisation and together with non-commercial systems such as R they started to influence the practice of analysis. However, we rather see the genetic components hitherto available as templates of future systems. The large number of genetic markers and a host of environmental covariates, which require more sophisticated modelling, will be the driving force for integrating genetic analysis into the usual analysis of epidemiological data. Their statement that *"flexible programs that allow sophisticated users to pursue novel models.and variations on existing models"* suggests itself as a flexible language in established systems. We would like to highlight R system here. Given the relative recent effort, the amount of contributed packages it contains is astonishing. The R system deserves particular attention since it is a result of collaborative work of many researchers, both academic and industrial, and portable to many computer systems. It is a flexible, comprehensive environment for statistical computing with an object-oriented programming language. It provides standard formats for data input, documentation and interface to general statistical package and databases. The interface to programs in C/C++/Fortran programs makes it possible to incorporate many stand-alone programs for genetic data analysis. The contributed packages are not necessarily limited to statistics but inclusive of other disciplines such as operations research, sociological methodologies and so on. Furthermore, standard datasets can be included as benchmarks for new statistical models. The benefit in educational use has repeatedly been shown [122, 123]. Seeing that the synthesis with the latest computing machinery in microarray analysis is already a catchword for biologists due to the development of Bioconductor (http://www.bioconductor.org), the outreach to many aspects of the analysis from CRAN is invaluable.

A major limitation of any review of this kind is the broad scope of the software systems makes it an impossible task. Fortunately, specific aspects of the software systems such as missing data analysis [124], multilevel modelling [125, 126], reliability [127], are available. Another limitation is the shortage of materials on other species [89, 128-130], although many packages in R are designed for them and the data fusion could be imminent. We have not covered packages for gene expression data in R or other platform such as Genstat [131] and S+ArrayAnalyzer; we expect such effort will be common in the future and software consolidation would be achieved. We only show the feasibility to use databases from these software systems, but the integration of genetic databases themselves remains to be achieved (e.g. International Workshop Integrative Bioinformatics http://www.rothamsted.bbsrc.ac.uk/bab/conf/ibiof/). Our focus on the overall picture of genetic analysis of complex traits and more generally genetic epidemiology, as well as the computing environment, by no means a sacrifice of the balance of analytic strategies and development of computing tools. It is reminiscent of a statement by Morton [39] *"As genetic epidemiology progresses it will reflect changes in population genetics...A new balance is being struck in which the study of contemporary populations eclipses evolutionary aspects, and mathematical biology increasingly takes its problems from the expansion of genetic epidemiology"* and that *"The cost of training a generation of researchers to value methods more than hypotheses has yet to be measured... Unless the trend toward directed research is reversed, genetic epidemiology will not be the only victim."* At the very time when genetic epidemiology is attracting many investigators and spawning a variety of analytical methods, the demand for inter-disciplinary synthesis is more important than ever.

Practically, the software consolidation and integrated genetic analysis entails to direct attention to established software systems as outlined here, with respect to aspects such as database capability, graphics, programming ability, Internet functionality, reliable and comprehensive numerical routines, documentation, use of available codes and availability. Another important aspect is to test using benchmark problems. The overlapping functions in these packages make it possible, e.g. the R package hapassoc and Stata with probability weight. Indeed, the software systems considered here have all been using example data to illustrate their analytical procedures concerning general statistics. Currently, there is a lack of linkage, e.g. identitiy-by-descent, routines available on these platforms, along with systematic power analysis.

To conclude, the wide availability of genomic data, the need of powerful statistical and computational tools, the limitation of individual researchers call for coordinated endeavours and integration with general computing environment are inevitable. This is a model that largely mirrors the development of Linux system in general computing, and R is mostly in line with this. As an alternative to other commercial systems, R offers the availability, variety and possibility for doing so. From our experience, this amount of effort would be less than most statisticians had previously anticipated. While some time is required for fully integrated analyses of most of the available genetic data, there will be exciting development to come and R is expected to be an important platform for advance.

ACKNOWLEDGEMENTS

We would like to thank the Editor for an invitation to contribute this review, Dr Andres Ruiz-Linares for the trio data, Prof Abbas Parsian for careful reading of the manuscript. JHZ is supported by MRC. QT wishes to acknowledge support from NIA grant NIA-P01-AG08761 and 'The MicroArray Center' project under the Biotechnological Research Program financed by the Danish Research Agency.

REFERENCES

[1] Gusella JF, Wexler NS, Conneally PM, *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **1983**; 306: 234-38.

[2] Tsui LC, Buchwald M, Barker D, *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **1985**; 230: 1054-57.

[3] Hall JM, Lee MK, Newman B, *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **1990**; 250: 1684-89.

[4] Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. *Science* **2002**; 298: 2345-49.

[5] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **2003**; 33 Suppl: 228-37.

[6] The International HapMap Project. *Nature* **2003**; 426: 789-96.

[7] Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. *Nature* **2005**; 437: 1299-320.

[8] Huttley GA, Smith MW, Carrington M, O'Brien SJ. A scan for linkage disequilibrium across the human genome. *Genetics* **1999**; 152: 1711-22.

[9] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **2004**; 74: 106-20.

[10] Hinds DA, Stuve LL, Nilsen GB, *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **2005**; 307: 1072-79.

[11] Wang DG, Fan JB, Siao CJ, *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **1998**; 280: 1077-82.

[12] Klein RJ, Zeiss C, Chew EY, *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **2005**; 308: 385-89.

[13] Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **2001**; 17: 481-85.

[14] Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* **2004**; 429: 475-77.

[15] Thomas DC, Haile RW, Duggan D. Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet* **2005**; 77: 337-45.

[16] Dudbridge F. A survey of current software for linkage analysis. *Hum Genomics* **2003**; 1: 63-65.

[17] Knight J. A survey of current software for genetic power calculations. *Hum Genomics* **2004**; 1: 225-27.

[18] Weale ME. A survey of current software for haplotype phase inference. **2004**; 1: 141-44.

[19] Salem RM, Wessel J, Schork NJ. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum Genomics* **2005**; 2: 39-66.

[20] Almasy L, Warren DM. Software for quantitative trait analysis. *Hum Genomics* **2005**; 2: 191-95.

[21] Guo S-W, Lange K. Genetic mapping of complex traits: Promises, problems, and prospects. *Theor Pop Biol* **2000**; 57: 1-11.

[22] Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* **1994**; 265: 2037-48.

[23] Olson JM, Witte JS, Elston RC. Genetic mapping of complex traits. *Stat Med* **1999**; 18: 2961-81.

[24] Risch NJ. Searching for genetic determinants in the new millennium. *Nature* **2000**; 405: 847-56.

[25] Zhao H. Family-based association studies. *Stat Meth Med Res* **2000**; 9: 563-87.

[26] Shih M-C, Whittemore AS. Allele-sharing among affected relatives: non-parametric methods for identifying genes. Stat Meth Med Res **2001**; 10: 27-55.

[27] Schaid DJ. Evaluating associations of haplotypes with traits. *Genet Epidemiol* **2004**; 27: 348-64.

[28] Thomas DC. Statistical Methods in Genetic Epidemiology Oxford, Oxford University Press 2004

[29] Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics* **2005**; 61: 899-911; discussion 11-41.

[30] Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussions). *J Am Stat Assoc* **2006**; 101: 89-118.

[31] Zhao JH, Tan Q. Integrated analysis of genetic data with R. *Hum Genomics* **2006**; 2: 258-65.

[32] Schork NJ. Genetics of complex disease: approaches, problems, and solutions. *Am J Respir Crit Care Med* **1997**; 156: S103-S09.

[33] Guo S-W. Genetic mapping of complex traits: promises, problems, and prospects. *Theoretical Population Biology* **2000**; 57: 1-11.

[34] Risch N, Merikangas K. The future of genetic studies of complex human diseases. **1996**; 273: 1516-17.

[35] Morton NE. Outline of Genetic Epidemiology Basel, Karger. **1982**.

[36] Morton NE, Rao DC, Lalouel J-M. Methods in Genetic Epidemiology, Karger. **1983**.

[37] Sham PC. Genetic Epidemiology. *Br Med Bull* **1996**; 52: 408-33.

[38] Khoury M, Beaty T, Cohen B. Fundamentals of Genetic Epidemiology, Oxford University Press. 1993

[39] Morton NE. Genetic epidemiology. *Annu Rev Genet* **1993**; 27: 523-38.

[40] Morton NE. Genetic epidemiology. *Ann Hum Genet* **1997**; 61 (Pt 1): 1-13.

[41] Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* **2005**; 366: 941-51.

[42] Jewell NP. Statistics for Epidemiology, ed. Chatfield C, Tanner M, Zidek J, Chapman & Hall/CRC **2004**.

[43] Rao DC, Morton NE, Yee S. Analysis of family resemblance. II. A linear model for familial correlation. *Am J Hum Genet* 1974; 26: 331-59.

[44] MacLean CJ, Morton NE, Elston RC, Yee S. Skewness in commingled distributions. *Biometrics* **1976**; 32: 695-99.

[45] Morton N, MacLean C. Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *Am J Hum Genet* 1974; 27: 365-84.

[46] Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* **1955**; 7: 277-318.

[47] Ott J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* **1974**; 26: 588-97.

[48] Lathrop G, Lalouel J, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* 1984; 81: 3443-46.

[49] Lander ES, Green P. Construction of multilocus genetic maps in humans. *Proc Nat Acad Sci USA* **1987**; 84: 2363-67.

[50] Self SG, Longton G, Kopecky KJ, Liang K-Y. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* **1991**; 47: 55-61.

[51] Jorde LB. Invited Editorial: linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* **1995**; 56: 11-14.

[52] Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res* **2000**; 10: 1435-44.

[53] Dawn Teare M, Barrett JH. Genetic linkage studies. *Lancet* **2005**; 366: 1036-44.

[54] Cordell HJ, Clayton DG. Genetic association studies. *Lancet* **2005**; 366: 1121-31.

[55] Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **2005**; 366: 1223-34.

[56] Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* **2005**; 366: 1315-23.

[57] Hopper JL, Bishop DT, Easton DF. Population-based family studies in genetic epidemiology. *Lancet* **2005**; 366: 1397-406.

[58] Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **2005**; 366: 1484-98.

[59] Gambaro G, Anglani F, D'Angelo A. Association studies of genetic polymorphisms and complex disease. *Lancet* **2000**; 355: 308-11.

[60] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* **2002**; 4: 45-61.

[61] Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **2002**; 3: 391-97.

[62] Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* **2004**; 429: 446-52.

[63] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **2005**; 6: 95-108.

[64] Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **2005**; 6: 109-18.

[65] Tan Q, Yashin AI, Christensen K, *et al.* Multidisciplinary approaches in genetic studies of human aging and longevity. *Curr Genomics* **2004**; 5: 409-16.

[66] Beeton M, Pearson K. Data for the problem of evolution in man, II, A first study of the inheritance of longevity and the selective death rate in man. *Proc R Soc London* **1899**; 65: 290-305.

[67] Robine JM, Allard M. The oldest human. *Science* **1998**; 279: 1834-5.

[68] Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, Vaupel JW. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. *Hum Genet* **1996**; 97: 319-23.

[69] De Benedictis G, Tan Q, Jeune B, Christensen K, Ukraintseva SV, Bonafe M, Franceschi C, Vaupel JW, Yashin AI. Recent advances in human gene-longevity association studies. *Mech Ageing Dev* **2001**; 122: 909-20.

[70] Puca AA, Daly MJ, Brewster SJ, Matise TC, Barrett J, Shea-Drinkwater M, Kang S, Joyce E, Nicoli J, Benson E, Kunkel LM, Perls T. A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. *Proc Natl Acad Sci USA* **2001**; 98: 10505-08.

[71] Smith JD. Apolipoproteins and aging: emerging mechanisms. *Ageing Res Rev* **2002**; 1: 345-65.

[72] Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Hum Hered* **1971**; 21: 523-42.

[73] Morton N, Rao D, Lalouel J. *Methods in Genetic Epidemiology*, Karger **1983**.

[74] Lange K, Weeks DE, Boehnke M. Programs for Pedigree Analysis: MENDEL, FISHER, dGENE. *Genet Epidmiol* **1988**; 5: 471-72.

[75] Cottingham RWJ, Idury RM, Schaffer AA. Faster sequential genetic linkage computations. *Am J Hum Genet* **1993**; 53: 252-63.

[76] Irwin M, Cox N, Kong A. Sequential imputation for multilocus linkage analysis. Proc Nat Acad Sci USA **1994**; 91: 11684-88.

[77] O'Connell JR, Weeks DE. The VITESSE algorithm for rapid exact multilocus linkage analysis *via* genotype set-recording and fuzzy inheritance. *Nat Genet* **1995**; 11: 402-08.

[78] Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* **1996**; 58: 1323-37.

[79] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. *Am J Hum Genet* **1996**; 58: 1347-63.

[80] Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* **1997**; 61: 1179-88.

[81] Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **1997**; 61: 748-60.

[82] Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **1998**; 62: 1198-211.

[83] Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* **2000**; 25: 12-13.

[84] Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **2002**; 30: 97-101.

[85] Elston RC, Gray-McGuire C. A review of the 'Statistical Analysis for Genetic Epidemiology' (S.A.G.E.) software package. *Hum Genomics* **2004**; 1: 456-59.

[86] Stram DO. Software for tag single nucleotide polymorphism selection. *Hum Genomics* **2005**; 2: 144-51.

[87] Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **1972**; 2: 3-19.

[88] Hasstedt SJ. Pedigree Analysis Package. **2002**, University of Utah: Salt Lake City.

[89] Labate JA. Software for population genetic analyses of molecular marker data. *Crop Science* **2000**; 40: 1521-28.

[90] Lewis PO, Zaykin D. Genetic Data Analysis: Computer program for the analysis of allelic data. **1997**, http://hydrodictyon.eeb.uconn.edu.

[91] Schneider S, Kueffer JM, Roessli D, Excoffier L. Arlequin: a software for population genetic analysis. **1998**, http://anthro.unige.ch/arlequin/.

[92] Ihaka R, Gentleman R. *R:* a language for data analysis and graphics. *J Comp Graph Stat* **1996**; 5: 299-314.

[93] Knapp M. Reconstructing parental genotypes when testing for linkage in the presence of association. *Theor Pop Biol* **2001**; 60: 141-48.

[94] Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* **1988**; 75: 800-02.

[95] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B* **1995**; 57: 289-300.

[96] Horvath S, Laird NM. A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* **1998**; 63: 1886-97.

[97] Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **1998**; 62: 450-58.

[98] Dudbridge F. Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* **2003**; 25: 115-21.

[99] Hilbe JM. A review of Stata 9.0. *Am Stat* **2005**; 59: 335-48.

[100] Johnson GCL, Esposito L, Barratt BJ, *et al.* Haplotype tagging for the identification of common disease genes. *Nat Genet* **2001**; 29: 233 - 37.

[101] Becker RA, Chambers JM, Wilks AR. The New S Language CA, Wasworth & Brooks/Cole, Pacific Grove **1988**.

[102] Venables WN, Ripley BD. Modern Applied Statistics with S, Springer **2002**.

[103] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature* **2000**; 405: 827-36.

[104] Lake SL, Lyon H, Tantisira K, *et al.* Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* **2003**; 55: 56-65.

[105] Therneau T, Atkinson B. Some S-PLUS tools for analysis of genetic data. in S-PLUS International User Conference. **2001**.

[106] Therneau T. On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees. **2003**, http://mayoresearch.mayo.edu/.

[107] Andrade Md, Atkinson EJ, Lunde E, Amos CI, Chen J. Estimating genetic components of variance for quantitative traits in family studies using the MULTIC routines. **2006**, http://mayoresearch.mayo.edu.

[108] Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet* **2005**; 76: 780-93.

[109] Devlin B, Roeder K. Genomic control for association studies. *Biometrics* **1999**; 55: 997-1004.

[110] Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that mat be subjected to parental imprinting. *Am J Hum Genet* **1998**; 62: 969-78.

[111] Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol* **1998**; 148: 893-901.

[112] Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **2003**; 100: 9440-45.

[113] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci* **2003**; 18: 71-103.

[114] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **2002**; 12: 111-39.

[115] Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc* **2004**; 99: 96-104.

[116] Efron B. *Local false discovery rates*. **2005**.

[117] Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics* **2005**; 21: 2921-22.

[118] Tufto J, Raybould AF, Hindar K, Engen S. Analysis of genetic structure and dispersal patterns in a population of sea beet. *Genetics* **1998**; 149: 1975-85.

[119] Sun N, Zhao H. Genomic approaches in dissecting complex biological pathways. *Pharmacogenomics* **2004**; 5: 163-79.

[120] Carey VJ, Gentry J, Whalen E, Gentleman R. Network structures and algorithms in Bioconductor. *Bioinformatics* **2005**; 21: 135-36.

[121] Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* **2002**; 53: 79-91.

[122] Nolan D, Speed TP. Teaching statistics theory through application. *Am Statist* **1999**; 53: 370-75.

[123] Horton NJ, Brown ER, Qian LJ. Use of R as a toolbox for mathematical statistics exploration. *Am Stat* **2004**; 58: 343-57.

[124] Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat* **2001**; 55: 244-54.

[125] Zhou X-H, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. *Am Stat* **1999**; 53: 282-90.
[126] Centre for Multilevel Modelling. Software Reviews of Multilevel Analysis Packages. **2006**, http://www.mlwin.com.
[127] Altman M, McDonald MP. Choosing reliable statistical software. *Political Sci Politics* **2001**; 34: 681-87.
[128] Darvasi A. Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet* **1998**; 18: 19-24.
[129] Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **2002**; 3: 43-52.
[130] Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants - Salient statistical tools and considerations. *Crop Science* **2003**; 43: 1235-48.
[131] Payne RW, Arnold GM. Genstat Release 6.1 Reference Manual - Part 3: Procedure Library PL14 Oxford, VSN International **2002**.

Received: November 9, 2005 Revised: March 2, 2006 Accepted: March 9, 2006