BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 10 2007, pages 1294–1296
doi:10.1093/bioinformatics/btm108

Genetics and population analysis

GenABEL: an R library for genome-wide association analysis

Yurii S. Aulchenko[1,*], Stephan Ripke[2], Aaron Isaacs[1] and Cornelia M. van Duijn[1]

[1]Department of Epidemiology and Biostatistics, Erasmus MC Rotterdam, Postbus 2040, 3000 CA Rotterdam, The Netherlands and [2]Statistical Genetics Group, Max-Planck-Institute of Psychiatry, Kraepelinstr. 10, D-80804 Munich, Germany

Received on December 3, 2006; revised on February 14, 2007; accepted on March 13, 2007

Advance Access publication March 23, 2007

Associate Editor: Martin Bishop

ABSTRACT

Here we describe an R library for genome-wide association (GWA) analysis. It implements effective storage and handling of GWA data, fast procedures for genetic data quality control, testing of association of single nucleotide polymorphisms with binary or quantitative traits, visualization of results and also provides easy interfaces to standard statistical and graphical procedures implemented in base R and special R libraries for genetic analysis. We evaluated GenABEL using one simulated and two real data sets. We conclude that GenABEL enables the analysis of GWA data on desktop computers.

Availability: http://cran.r-project.org

Contact: i.aoultchenko@erasmusmc.nl

1 INTRODUCTION

Genome-wide association (GWA) analysis is a tool of choice for the identification of genes for complex traits. Effective storage, handling and analysis of GWA data represent a challenge to modern computational genetics. GWA studies generate large amounts of data: hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped in hundreds or thousands of patients and controls. Data on each SNP undergoes several types of analysis: characterization of frequency distribution, testing of Hardy–Weinberg equilibrium, analysis of association between single SNPs and haplotypes and different traits and so on. Because SNP genotypes in dense marker sets are correlated, significance testing in GWA analysis is preferably performed using computationally intensive permutation test procedures, further increasing the computational burden (Evans and Cardon, 2006).

Effective software making GWA analysis possible on desktop computers should meet the following criteria:

1. Facilitate effective data storage and manipulation.
2. Give access to wide range of statistical and graphical tools.
3. Implement fast procedures for specific GWA tests.

With these objectives in mind, we developed the GenABEL software, implemented as an R library. R is a free, open source language and environment for statistical analysis (http://www.r-project.org/). Building upon existing statistical analysis facilities allowed for rapid development of the package.

2 IMPLEMENTATION

2.1 Objective (1)

GWA data storage using standard R data types is ineffective. A SNP genotype for a single person may take four values (AA, AB, BB and missing). Two bits, therefore, are required to store these data. However, the standard R data types occupy 32 bits, leading to an overhead of 1500%, compared to the theoretical optimum. Use of the raw R data format, occupying eight bits, would still lead to 75% of RAM used inefficiently; moreover, this data type cannot be used directly in an analysis. We developed a new R data class, snp.data, which uses the optimal two bits to store information on a single SNP genotype. The standard R subsetting model was applied for this class, allowing retrieval of subsets of the data by SNP and study subject index, name or logical condition. Coercion to R integer and character and data types used by the ‘‘haplo.stats’’ (Schaid *et al*., 2002) and ‘‘genetics’’ libraries was implemented.

2.2 Objective (2)

R provides extensive statistical analysis and graphical facilities. This was one of the reasons why we implemented GenABEL as an R library. The function scan.glm and scan.glm.2D were developed to iteratively apply the standard R procedure glm (estimation of generalized linear models) to GWA data. The functions scan.haplo and scan.haplo.2D use the ‘‘haplo.stats’’ library to run sliding-window haplotype analysis and to evaluate the associations between a trait and haplotypes formed by all possible pairs of SNPs in a region. These functions are relatively slow and are aimed at the analysis of selected regions. In order to represent the objects generated by GenABEL graphically, new methods were designed for the generic R function ‘‘plot.’’

2.3 Objective (3)

Fast statistical genetic analysis procedures were implemented using ANSI standard of the C language and integrated into our

*To whom correspondence should be addressed.

library. These procedures facilitate data quality control and rapid single-SNP GWA analysis. The check.trait function provides summary statistics for phenotypic data and checks for outliers at a specified *P*-value or false discovery rate cut-off level. The function check.marker, based on summary.snp.data, allows the selection of a set of SNPs which pass user-specified criteria on call rate, redundancy, minimal marker allele frequency and deviation from Hardy–Weinberg equilibrium (using an exact test, Wigginton *et al.*, 2005). The functions ccfast and qtscore enable a fast GWA analysis for case-control data and quantitative traits. The functions emp.ccfast and emp.qtscore were developed to estimate empirical genome-wide significance.

3 EXAMPLE

We applied GenABEL for the analysis of one simulated and two real data sets. The first data set is distributed together with GenABEL. Using the MS program (Hudson, 2002), 833 SNPs covering a 2.5 MB region were simulated in 2500 people. We denote this data set as 2500 × 0.8 k. Two real data sets both used Affymetrix 250 K SNP arrays. The first included 197 (197 × 250 k) and the second included 500 people (500 × 250 k). All analyses were performed on a workstation with a 64-bit Intel Xeon 2.8 GHz processor, running SuSE Linux 9.2, using R v. 2.4.1. Analysis under Windows 2000 showed similar benchmark results.

Table 1 shows maximum resident memory size used by the package. It should be noted that most of the memory is occupied by the descriptive data (such as SNP names) and objects storing the results of analysis. For the 250 K GWA data, the maximum resident memory set size was 402 MB (set containing 500 people). A data set, which was obtained by quadruplicating every person in the 500 × 250 k set occupied a maximum of 1.24 GB. The memory occupied is roughly proportional to the number of subjects, though if the number of subjects increases *N* times, the RAM required increases by less than *N* times. Thus, GenABEL will facilitate analysis of GWA data on at least 2500 subjects on desktop computers (RAM 2 GB). From Table 1, it is clearly possible to run GWA and regional analyses in the course of a few minutes. Estimation of empirical genome-wide significance is one of the most laborious parts. Time for analysis grows proportionally to the product of the number of subjects, SNP tests and analysis replicas. Again, as is the case with RAM, with an *N*-fold increase of this product, time for computations increase slightly less than *N* times. Using GenABEL, it was possible to estimate empirical genome-wide significance using 500 permutations in a data set of 2000 people within 76 min.

Table 1. Characteristics of GenABEL v. 1.1–6

Characteristic	Data set		
	2500 × 0.8 k	197 × 250 k	500 × 250 k
Maximum resident memory usage, MB	49	205	402
Time for GWA analysis			
Loading the data, load.gwaa.data	<1 s	22 s	31 s
SNPs characterization, summary	<1 s	4 s	6 s
Case-control tests, ccfast	<1 s	2 s	4 s
Score tests, qtscore	<1 s	4 s	14 s
Genome-wide significance, emp.ccfast	14 s	3.5 min	5.5 min
Genome-wide significance, emp.qtscore	20 s	5 min	9 min
Time for analysis of a region of 21 SNPs			
2D interaction analysis, scan.glm.2D	40 s	18 s	21 s
2-SNPs haplotype analysis, scan.haplo	19 s	4 s	5 s
3-SNPs haplotype analysis, scan.haplo	29 s	4 s	6 s
2D haplotype analysis, scan.haplo.2D	3.5 min	35 s	52 s

Genome-wide significance was assessed using 100 randomly permuted samples.

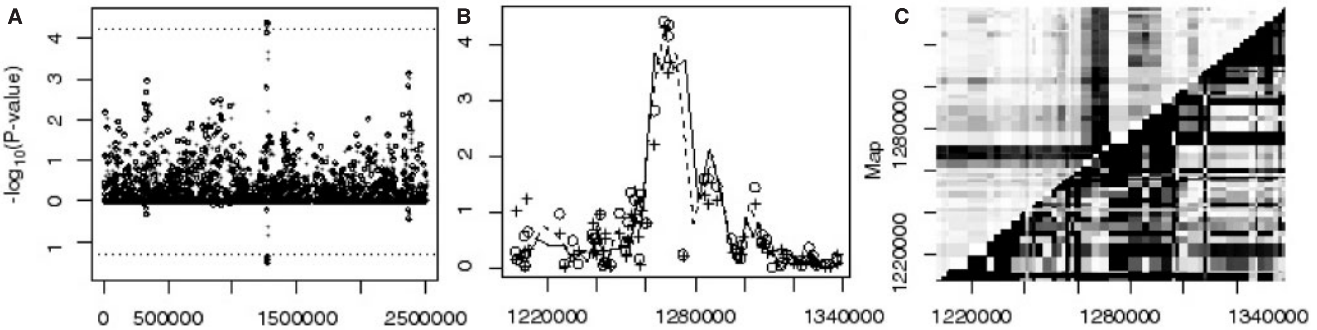

Fig. 1. Analysis of simulated data set. Region-wide analysis of single SNP association, using qtscore. (**A**) Nominal (above zero) and region-wise empirical (below zero) significance is presented as $-\log_{10}P$. Dotted lines correspond to experimentwise 5% significance (Bonferroni corrected above and empirical for below zero). Dots: allelic 1 d.f. test; crosses: genotypic 2 d.f. test. (**B**) Analysis of region surrounding SNPs showing highest significance. Dotted line: two-SNP sliding window haplotype analysis; solid line: three-SNP sliding window analysis. (**C**) Analysis of all pairs of SNPs in the region. Intensity corresponds to $-\log_{10}P$ from analysis of haplotype association (above diagonal) and D′ (below diagonal).

GenABEL facilitates not only GWA analysis, but also presentation of results. Figure 1 presents graphs generated in the analysis of the 2500 × 0.8 k set. In Figure 1A, the associations between the simulated quantitative trait and SNPs are shown for the whole region. Figure 1B and C presents results of more detailed analyses of the region surrounding the most significant association signal.

4 CONCLUSIONS

We developed the GenABEL package for GWA analysis, which implements effective GWA data storage and handling, fast procedures for genetic data quality control and analysis and interfaces to standard and specific R data types and functions. The package is available at http://cran.r-project.org.

ACKNOWLEDGEMENTS

We would like to thank Prof. L.Cardon, Prof. D.Clayton, Dr B.Müller-Myhsok and Dr M.Kayser for their valuable insights. This work was supported by the Netherlands Organization for Scientific Research (NWO-RFBR 047.016.009), the Centre for Medical Systems Biology (CMSB), the European Special Populations Research Network (FP6) and the Russian Foundation for Basic Research (RFBR). The work of Y.S.A was supported by a grant from ‘Stichting MS’.

Conflict of Interest: none declared.

REFERENCES

Evans,D.M. and Cardon,L.R. (2006) Genome-wide association: a promising start to a long race. *Trends Genet*., **22**, 350–354.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Schaid,D.J. *et al*. (2002) Score tests for association between traits and hap lotypes when linkage phase is ambiguous. *Am. J. Hum. Genet*., **70**, 425–434.

Wigginton,J.E. *et al*. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet*., **76**, 887–893.