BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 8 2012, pages 1168–1169
doi:10.1093/bioinformatics/bts086

Genome analysis

Advance Access publication February 17, 2012

PERMORY-MPI: a program for high-speed parallel permutation testing in genome-wide association studies

Volker Steiß[1,2], Thomas Letschert[2], Helmut Schäfer[1] and Roman Pahl[1,*]

[1]Institute of Medical Biometry and Epidemiology, Philipps-Universität Marburg, 35037 Marburg and [2]Institute of Software Architecture, Technische Hochschule Mittelhessen, 35390 Gießen Germany

Associate Editor: Alex Bateman

ABSTRACT

Summary: PERMORY is software for accelerated permutation testing of genome-wide association studies (GWAS). We have parallelized PERMORY using the Message-Passing Interface resulting in a nearly linear speedup. Furthermore, we added accelerated analysis of GWAS using quantitative phenotypes, and an accurate estimation of the effective number of independent tests.

Availability and implementation: Free download from http://permory.org

Contact: rpahl@staff.uni-marburg.de; volker.steiss@mni.thm.de

Received on November 9, 2011; revised on February 3, 2012; accepted on February 13, 2012

1 INTRODUCTION

A genome-wide association study (GWAS) nowadays involves the statistical analysis of millions of single nucleotide polymorphism (SNP) markers. For the calculation of adjusted *P*-values for genome-wide testing, the Bonferroni method is highly conservative, because it ignores the correlation of markers caused by linkage disequilibrium. In contrast, a permutation-based correction of *P*-values fully accounts for this correlation and provides the highest statistical power among the procedures controlling genome-wide type I error risk. While conventional permutation algorithms often required prohibitively long computation times for large datasets, the PERMORY algorithm of Pahl and Schäfer (2010) has made permutation testing feasible for large-scale GWAS analysis. In time, SNP chips with larger numbers of markers, such as Illumina® Human Omni with 5 Mio SNPs are available, and the next-generation sequencing techniques will open the way for full sequence association studies adding another 10 or 20 million of rare variants. We therefore have undertaken one step further and developed the first freely available parallel algorithm for permutation testing in GWAS. Based on the original PERMORY algorithm, the computation process now can be spread to multiple processors in a cluster using MPI (Message Passing Interface Forum, 2009). In addition, we have further developed the original algorithm, being restricted to the analysis of dichotomous phenotypes, now to handle quantitative phenotypes as well. Moreover, for each dataset, PERMORY now also provides an accurate estimate of the effective number of independent tests (Cheverud, 2001).

*To whom correspondence should be addressed.

2 DESCRIPTION OF PERMORY-MPI

PERMORY-MPI[1] is an open source program written in C++ using the C++ library Boost MPI. It can be run on every computer cluster where an MPI implementation is installed, for example Open MPI version 1.2.7 rc2 or higher (http://www.open-mpi.org/). The data files must be available at the specified path on each node PERMORY-MPI should run at. A network file system is useful in this situation. If a network file system is not available, or the throughput of the network file system is insufficient, the data files have to be copied to each node before submitting the job to the MPI cluster.

The program supports various input data formats (PLINK, Purcell *et al.* 2007; PRESTO, Browning 2008; SLIDE, Han *et al.* 2009). The format is detected automatically. Files of different formats can be used in combination. Marker genotypes may be coded as 0, 1 and 2 for the copy number of a specific allele, or in allelic representation as 'AC' and 'GT'. Phenotypic values can be real numbers coding for quantitative traits, or 0/1 or 1/2 coding for unaffected/affected, respectively. The type of trait, quantitative or dichotomous, is determined automatically by the software.

Regarding dichotomous traits, PERMORY is using Cochran–Armitage's trend test, where, for each marker, a test statistic is derived from the contingency table that contains the genotype frequencies for cases and controls (Pahl and Schäfer, 2010). The handling of quantitative traits is a novel feature, which requires a different type of statistic. PERMORY-MPI uses the test statistic defined by Lin (2006)

$$T_j = \frac{\left(\sum_{i=1}^{n} U_{ji}\right)^2}{\sum_{i=1}^{n} U_{ji}^2}, \quad U_{ji} = \sum_{i=1}^{n} (Y_i - \mu_y)(X_{ji} - \mu_j)$$

where Y_i is the phenotypic value of the i-th subject, X_{ji} is the genotype score for the j-th marker of the i-th subject, and μ_y and μ_j are the population means of Y_i and X_{ji}. To maintain the computational speed of the original algorithm, the two most important acceleration methods, genotype indexing with transposed permutations and reconstruction memoization (Pahl and Schäfer 2010), were further developed to be applicable to the above statistic (Steiß, 2011).

For the mentioned statistics the asymptotic distributions are known so that a raw *P*-value, say p_j, can be derived for each of the $j = 1, \ldots, M$ markers. To correct p_j for the multiplicity of tested hypotheses, the distribution under the null of no marker being

[1]While the official name of the software has remained PERMORY, we use the term PERMORY-MPI throughout the article to emphasize new features of the software.

associated with the trait is simulated by permuting the phenotype values (or trait labels). In particular, the genome-wide adjusted P-value for marker j is calculated using a single step max-t procedure (Westfall and Young, 1993). Assuming a set of, say, r permutations, $k=1,\ldots,r$, of the phenotype vector, we set

$$p_j^* = \left(1 + \#\text{ of } k\text{'s for which } \tilde{T}_{\max}^k > T_j\right)/(r+1)$$

where T_j is the value of the test statistic for marker j in the original sample, and $\tilde{T}_{\max}^k$ is the maximal test statistic over all markers under the k-th permutation. With a number of P available processors, PERMORY-MPI delegates the calculation of r/P maximal values $T_{\max}^k$ to each processor. These values are then merged on the central process, where the adjusted P-values are calculated. The final output consists of a sorted list of a user-specified number of top markers, providing two-sided raw (i. e., unadjusted) and genome-wide adjusted P-values for every marker.

Permutation tests are considered the gold standard for multiple testing correction in GWAS. One important alternative approach is to estimate the effective number of independent tests M_{eff} (Cheverud, 2001) for a set of M markers, and use this number for an improved Bonferroni correction: $p_j^* = p_j \cdot M_{\text{eff}}$. The more accurate M_{eff} can be estimated, the more precise will be the P-value adjustment. Since by simple algebra $M_{\text{eff}} = p_j^*/p_j$, the adjusted P-value obtained from the permutation test yields a direct estimate of M_{eff}. PERMORY-MPI now calculates M_{eff} for the top marker, which, assuming a sufficient number of permutations, provides a more accurate estimate than available 'indirect' approaches (e. g. Cheverud 2001; Gao *et al.* 2008; Li and Ji 2005).

Any method or test realizable by the presented permutation approach can be incorporated into the software. The implementation of additional methods is supported by the object oriented software design of PERMORY. In particular, stratified testing and the inclusion of covariates are planned to be implemented in a next version.

3 RESULTS

We investigated the runtime behavior of PERMORY-MPI using genotype data from the WTCCC2 study (The Wellcome Trust Case Control Consortium, 2007) with 1 115 428 SNPs in 5667 individuals, randomly divided into 2834 cases and 2833 controls. The benchmarks resulted in a nearly linear speedup (Fig. 1). A conservative extrapolation of the results of Pahl and Schäfer (2010) would result in runtimes of more than 250 h with PRESTO and of more than 15 000 h with PLINK for the present dataset. We also compared PLINK and PERMORY for quantitative phenotypes, using the genotype data of the WTCCC2 study in Chromosome 22 with 18 649 SNPs combined with randomly generated quantitative phenotypes following a Gaussian distribution. With 10 000 permutations, runtimes were 0.27 h with single process PERMORY, and 0.04 h using 10 parallel processes, as compared to 52.38 h with PLINK.

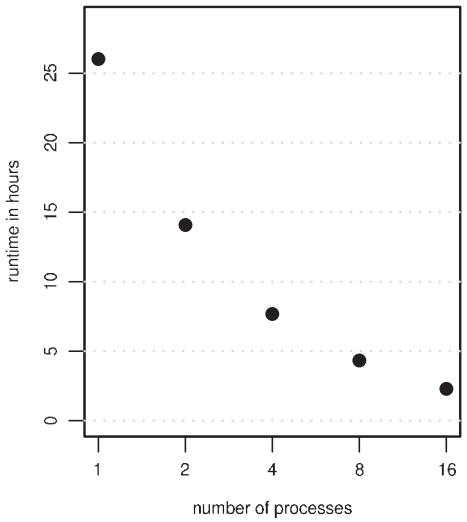

Fig. 1. Runtimes of PERMORY and PERMORY-MPI using different numbers of processes analysing data from the WTCCC2 study with 1 115 428 SNPs, 5667 individuals and 100 000 permutations, showing a nearly linear speedup.

ACKNOWLEDGEMENT

We thank the reviewers for their valuable comments.

Funding: This work was funded by von Behring-Röntgen-Stiftung, grant no. 57–0048.

Conflict of Interest: none declared.

REFERENCES

Browning,B.L. (2008) PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted *P*-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*, **9**, 309.

Cheverud,J.M. (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87**, 52–58.

Gao,X. *et al.* (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol*, **32**, 361–369.

Han,B. *et al.* (2009) Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*., **5**, e1000456.

Li,J. and Ji,L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*, **95**, 221–227.

Lin,D.Y. (2006) Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.*, **78**, 505–509.

Message Passing Interface Forum (2009) MPI: a message-passing interface standard version 2.2. *Technical report*, Message Passing Interface Forum, University of Tennessee, Knoxville, Tennessee.

Pahl,R. and Schäfer,H. (2010). Permory: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics*, **26**, 2093–2100.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Steiß, V. (2011) Extending the PERMORY-algorithm: parallelization and handling of quantitative phenotypes. Bachelor's Thesis, Faculty of Mathematics, Natural Sciences and Information Technology, Technische Hochschule Mittelhessen – University of Applied Sciences, Gießen.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons, New York.