Bioinformatics, 31(9), 2015, 1460–1462
doi: 10.1093/bioinformatics/btu840
Advance Access Publication Date: 21 December 2014
Applications Note

Genome analysis

PBOOST: a GPU-based tool for parallel permutation tests in genome-wide association studies

Guangyuan Yang[1,*], Wei Jiang[1], Qiang Yang[2] and Weichuan Yu[1,*]

[1]Laboratory of Bioinformatics and Computational Biology, Department of Electronic and Computer Engineering and [2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 8, 2014; revised on December 9, 2014; accepted on December 16, 2014

Abstract

Motivation: The importance of testing associations allowing for interactions has been demonstrated by Marchini *et al*. (2005). A fast method detecting associations allowing for interactions has been proposed by Wan *et al*. (2010a). The method is based on likelihood ratio test with the assumption that the statistic follows the χ^2 distribution. Many single nucleotide polymorphism (SNP) pairs with significant associations allowing for interactions have been detected using their method. However, the assumption of χ^2 test requires the expected values in each cell of the contingency table to be at least five. This assumption is violated in some identified SNP pairs. In this case, likelihood ratio test may not be applicable any more. Permutation test is an ideal approach to checking the *P*-values calculated in likelihood ratio test because of its non-parametric nature. The *P*-values of SNP pairs having significant associations with disease are always extremely small. Thus, we need a huge number of permutations to achieve correspondingly high resolution for the *P*-values. In order to investigate whether the *P*-values from likelihood ratio tests are reliable, a fast permutation tool to accomplish large number of permutations is desirable.

Results: We developed a permutation tool named PBOOST. It is based on GPU with highly reliable *P*-value estimation. By using simulation data, we found that the *P*-values from likelihood ratio tests will have relative error of >100% when 50% cells in the contingency table have expected count less than five or when there is zero expected count in any of the contingency table cells. In terms of speed, PBOOST completed 10^7 permutations for a single SNP pair from the Wellcome Trust Case Control Consortium (WTCCC) genome data (Wellcome Trust Case Control Consortium, 2007) within 1 min on a single Nvidia Tesla M2090 device, while it took 60 min in a single CPU Intel Xeon E5-2650 to finish the same task. More importantly, when simultaneously testing 256 SNP pairs for 10^7 permutations, our tool took only 5 min, while the CPU program took 10 h. By permuting on a GPU cluster consisting of 40 nodes, we completed 10^{12} permutations for all 280 SNP pairs reported with *P*-values smaller than 1.6×10^{-12} in the WTCCC datasets in 1 week.

Availability and implementation: The source code and sample data are available at http://bioinformatics.ust.hk/PBOOST.zip.

Contact: gyang@ust.hk; eeyu@ust.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In genome wide association studies (GWAS), single nucleotide polymorphism (SNP) pairs have significant associations with diseases via the combination of their main effects and interactions. This effect is referred to as associations allowing for interactions (Marchini *et al.*, 2005). Wan et al. (2010a) proposed an efficient likelihood ratio test method and reported 280 SNP pairs having significant associations allowing for interactions. This method is based on χ^2 test and requires adequate expected count in each cell of the contingency table. Otherwise, the statistic may not follow the χ^2 distribution, and the likelihood ratio test is not applicable. Cochran (1952) suggested a rule of thumb that at least 80% cells should have expected counts larger than or equal to five. There are also discussions claiming that there should be no cell with zero expected count (Conover, 1999). In GWAS, it is common to see SNP pairs whose contingency tables violate this rule. Thus, we like to examine the likelihood ratio test results with the permutation test method.

In a permutation test, the *P*-value is calculated as $(N_0+1)/(N+1)$, where N_0 is the number of permutations with statistics at least as extreme as the one of the original data, and N is the total number of permutations. The resolution of the *P*-value is $1/(N+1)$. Clearly, a large number of permutations are required to estimate a very small *P*-value. In GWAS, we always need the *P*-value to be extremely small due to the multiple testing issue. In the WTCCC datasets, around 3.5×10^5 SNPs remain after initial quality check. The total number of SNP pairs reaches 6.1×10^{10}. The *P*-value needs to be smaller than 1.6×10^{-12} to denote significance when multiple testing correction (such as Bonferroni correction) is applied.

The development of graphics processing units (GPUs) enables fast implementation of permutation. Here, we implemented a GPU-based permutation test tool. Using this tool, we investigated the eligibility condition of the likelihood ratio test. We also examined SNP pairs with associations allowing for interactions reported by likelihood ratio test.

2 Methods

In the label permuting stage, we adopted the shuffle algorithm (Fisher and Yates, 1949). To parallelize the permutation test, we decided to generate independent random sequences parallelly using Mersenne twister (Matsumoto and Nishimura, 1998) with a very long period of $2^{19937}-1$.

In the statistic calculation stage, we revised the bitset representation of the genotype data proposed by Wan *et al.* (2010b). Firstly, both the genotypes and the labels are represented by bit vectors. We can permute on the label vector efficiently without changing the genotypes. Then in the statistic calculation stage, the contingency tables can be collected by simple logic operations. Secondly, the genotype data could be stored in the shared memory which has small latency and less memory access restriction. Thirdly, the marginal distributions of the observations are invariant in permutation (Pahl and Schäfer, 2010). Thus, we only need to count eight cells of the contingency table. The others were derived by the marginal distributions. This saved ~56% Boolean operations.

Moreover, we designed the staggered memory layout to optimize the memory bandwidth. In this memory layout, the memory accesses of the threads running on a multiprocessor were coalesced. It reduced ~40% permuting time and 80% statistic calculation time. Please refer to the Supplementary document for implementation details.

In a typical graphics card, such as Nvidia Tesla M2090, there are 512 processing cores. By designing the memory access pattern, we achieved ~100 times speedup in permutation. Furthermore, all SNP pairs could share the same label vector. We can use permutation to test all the reported significant SNP pairs concurrently with only a little extra computation time.

In order to investigate the eligibility condition of likelihood ratio test, we simulated datasets consisting of SNP pairs with contingency tables having 0, 1, 2, ..., 10 cells whose expected count is less than five and also simulated 2, 4, 6 cells whose expected count equals to zero. There are 100 pairs and 1000 balanced samples in each dataset and their likelihood ratio test based *P*-values are between 10^{-9} and 10^{-7}. We recalculated these *P*-values using 10^{10} iterations of permutation test. The average relative errors of *P*-values we calculated as $\frac{1}{N}\sum_{i=1}^{N}|logp_i^{(p)} - logp_i^{(l)}|$, where $p_i^{(p)}$ and $p_i^{(l)}$ are *P*-values of permutation test and likelihood ratio test, respectively. Figure 1 illustrates the relative error of *P*-value under different contingency table conditions.

Generally, the average error increased as the contingency table deteriorated. We did *t*-test for $\{p_i^{(p)}\}$ and $\{p_i^{(l)}\}$ with the null hypothesis that there was no significant difference between them. We got $p_{t-test} = 0.0271$ when there are two cells with zero expected count, and $p_{t-test} = 0.0114$ when there are 50% cells with expected count less than five. Thus, the null hypothesis was rejected under 0.05 significance level when there was zero expected count in any of the contingency table cells or at least 50% cells have expected count less than five.

We also did permutation tests on the SNP pairs with associations allowing for interactions in the WTCCC datasets. We checked all the identified pairs reported by Wan *et al.* (2010a). There were three pairs whose contingency tables have two cells with expected count equals to zero. In these three pairs, we found two pairs with large relative errors between the *P*-values from likelihood ratio tests and those from permutation tests. The permutation test results indicated that these two pairs failed to pass the significance threshold. The third pair had $N_0 = 0$ after 10^{12} permutations. Its *P*-value from permutation test was smaller than 10^{-12}. We could not calculate its relative error due to the limited *P*-value resolution in 10^{12} permutations, but this limited resolution of 10^{-12} indeed helped us verify that the corresponding association allowing for interaction was statistically significant even after Bonferroni correction. (Please refer to the Supplementary for the detailed results). In this sense, our tool is indeed useful to check the *P*-values from likelihood ratio tests.

There were another 10 pairs having >20% cells in their contingency tables with expected count less than five. For all these pairs, there was no significant relative error in their *P*-values under the resolution of 10^{-12}. This result is consistent with the experiment conclusion on simulation datasets.

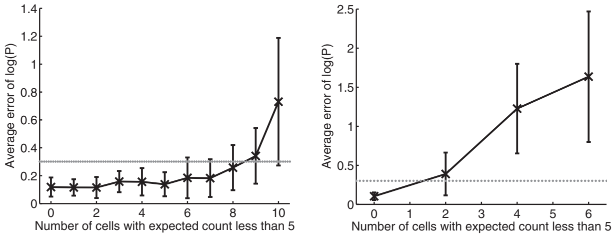

Fig. 1. Relative error of *P*-values under different conditions of contingency tables. The dashed line is the 100% relative error. When there are 50% cells with expected count less than five or when there are cells with zero expected count, the average relative error is >100%

3 Discussion

PBOOST makes it possible to test associations allowing for interactions using permutation. The empirical rule requiring that no >20% cells having expected count less than five is conservative. It can be relaxed to that <50% cells having expected count less than five and no cell having zero expected count. In this article, we only focused on the identified SNP pairs. In fact, there are a lot more non-significant SNP pairs (3.7×10^9 in Type-I diabetes dataset) which also violate the above rule. Some of them might be significant. Testing these SNP pairs remains an open challenge.

Acknowledgement

This work is partially supported by theme-based research project T12-402/13N from the Research Grant Council of the Hong Kong S.A.R. Government of China.

Conflict of interest: none declared.

References

Cochran,W.G. (1952) The χ^2 test of goodness of fit. *Ann. Math. Stat.*, **23**, 315–345.

Conover,W.J. (1999) *Practical Nonparametric Statistics*. John Wiley and Sons, New York.

Fisher,R.A. and Yates,F. (1949) *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd edn. Oliver & Boyd: London.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Matsumoto,M. and Nishimura,T. (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.*, **8**, 3–30.

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

Wan,X. *et al.* (2010a) Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics*, **26**, 2517–2525.

Wan,X. *et al.* (2010b) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

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.