BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 5 2010, pages 694–695
doi:10.1093/bioinformatics/btq009

Genetics and population analysis

Advance Access publication January 16, 2010

Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS

Casey S. Greene[1,†], Nicholas A. Sinnott-Armstrong[1,†], Daniel S. Himmelstein[1], Paul J. Park[2], Jason H. Moore[1,*] and Brent T. Harris[2]

[1]Department of Genetics and [2]Department of Pathology, Dartmouth Medical School, Lebanon, NH 03756, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Epistasis, the presence of gene–gene interactions, has been hypothesized to be at the root of many common human diseases, but current genome-wide association studies largely ignore its role. Multifactor dimensionality reduction (MDR) is a powerful model-free method for detecting epistatic relationships between genes, but computational costs have made its application to genome-wide data difficult. Graphics processing units (GPUs), the hardware responsible for rendering computer games, are powerful parallel processors. Using GPUs to run MDR on a genome-wide dataset allows for statistically rigorous testing of epistasis.

Results: The implementation of MDR for GPUs (MDRGPU) includes core features of the widely used Java software package, MDR. This GPU implementation allows for large-scale analysis of epistasis at a dramatically lower cost than the standard CPU-based implementations. As a proof-of-concept, we applied this software to a genome-wide study of sporadic amyotrophic lateral sclerosis (ALS). We discovered a statistically significant two-SNP classifier and subsequently replicated the significance of these two SNPs in an independent study of ALS. MDRGPU makes the large-scale analysis of epistasis tractable and opens the door to statistically rigorous testing of interactions in genome-wide datasets.

Availability: MDRGPU is open source and available free of charge from http://www.sourceforge.net/projects/mdr.

Contact: jason.h.moore@dartmouth.edu

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

Received on October 16, 2009; revised on January 7, 2010; accepted on January 8, 2010

1 INTRODUCTION

Genome-wide association studies hold promise for the discovery of the genetic factors that underlie common human diseases (Hirschhorn and Daly, 2005; Wang *et al.*, 2005). Unfortunately this promise has largely not been realized (Shriner *et al.*, 2007; Williams *et al.*, 2007). It is thought that this failure could be due to epistasis, the role of gene–gene interactions, which has commonly been ignored in these studies. Powerful and model-free methods such as multifactor dimensionality reduction (MDR) have been developed (Ritchie *et al.*, 2001), but an exhaustive examination of even pair-wise interactions in a 550 000 SNP dataset would require the analysis of 1.5×10^{11} combinations. While an analysis of this scale is approachable with modern cluster computing, an analysis that includes permutation testing to assess the statistical significance of results remains infeasible with CPU-based approaches.

Rendering photo-realistic video games in real time is also computationally difficult. For video game graphics, specific hardware (the graphics processing unit or GPU) has been developed. The GPU is a massively parallel computing platform that can be adapted to some scientific tasks. We have previously shown that MDR is one of these tasks (Sinnott-Armstrong *et al.*, 2009). Here we provide software which makes practical the analysis of epistasis in genome-wide data through the use of GPUs and demonstrate its application to a genome-wide analysis of epistasis of sporadic amyotrophic lateral sclerosis (ALS).

2 METHODS

MDRGPU, a software tool capable of analyzing genome-wide data, is a Python implementation of MDR, which uses the PyCUDA library to run MDR on GPUs. MDRGPU 1.0 supports balanced accuracy, large datasets, execution across an arbitrary number of GPUs, permutation testing and the analysis of high-order interactions. It runs on GPUs which support CUDA (i.e. the NVIDIA GeForce 8800 series and higher). Parallel execution of one realization across multiple GPUs is supported with the pp library for Python. MDRGPU provides a command-line interface for scripted analysis.

The GPU architecture has various memory spaces available. MDRGPU uses the constant cache, global memory, shared memory and registers. Shared memory is used to store the intermediate case and control counts for each attribute combination and to store the number of true and false positives and negatives. The global memory is accessed directly to fetch attributes. The constant cache is used in MDRGPU to store the case–control status. Dataset sizes of greater than 65 536 attributes require splitting which is handled seamlessly by MDRGPU. This splitting does not cause linear slowdown; there is simply more overhead of launching, so datasets with large numbers of instances see less of a performance reduction than datasets with few instances. The largest number of addressable attributes is 4 billion requiring 4 GB RAM per instance. In order for the case–control status to be held in constant memory, there can be at most 16 384 instances.

Our proof of concept analysis was performed on three GPU workstations (detailed in Supplementary Material S1). These systems contain three GeForce 295 cards, each of which contains two GPUs. For the first stage of this analysis, we used an ALS dataset from Schymick *et al.* (2007) as our detection dataset. This dataset was obtained from QUEUE at Coriell, but has since been moved to dbGaP. It contains 276 individuals with sporadic ALS and 271 control individuals. These individuals are genotyped at 555 352 SNPs using the Illumina Infinium II HumanHap550 SNP chip. We processed this dataset by removing SNPs with a minor allele frequency <0.2 or those

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

in which >10% of values were missing for either cases or controls. We further used Haploview's tagSNP algorithm (Barrett *et al.*, 2005) to select representative SNPs from groups of correlated SNPs ($r > 0.8$). After this, 210 382 SNPs remained and were used in the analysis. For the replication stage, we used a dataset of Irish individuals containing of 221 sporadic ALS patients and 211 controls described in Cronin *et al.* (2008).

We used MDRGPU to perform a two-way analysis across the entire detection dataset. We selected the SNP combination with the best balanced accuracy measure. We then permuted the dataset 1000 times while repeating this analysis. We measured the accuracy of the best pair in each permuted dataset. We then used the 50th best accuracy obtained from these permuted datasets as our significance cutoff. This permutation test yields an experiment-wise α of 0.05. A pair of SNPs with a significant association in the detection phase was tested in the replication dataset. In this phase, the two detected SNPs were selected from the dataset and MDR was used to evaluate only this pair. A permutation test was performed here using MDR on only these two SNPs, and an α of 0.05 was used to assess significance.

3 RESULTS

Our three GPU systems completed an analysis of pairwise interactions in a single permutation approximately every 6 min. The time to analyze the dataset itself for pairwise interactions is the same as the time required for one permutation. One thousand permutations were used to assess statistical significance which required ~100 h. The time to analyze the same dataset on a cluster with 200 AMD Opteron 2384 (2.7 GHz) CPU cores was just over 1 h without permutation testing and thus a CPU-based permutation test was considered infeasible as the estimated time required on 200 CPU cores was >40 days.

In the proof-of-concept analysis, the highest accuracy combination in our dataset was SNPs rs4363506 and rs6014848 with a balanced accuracy of 0.6551. In our permutation test, this accuracy was statistically significant ($P < 0.048$). In the replication dataset this pair had a balanced accuracy of 0.5821. Permutation testing the replication dataset showed that this result was also statistically significant ($P < 0.021$). Therefore, not only have we discovered a statistically significant pair of SNPs using an experiment-wise α of 0.05, but we have replicated the significant relationship in an independent dataset. Here is evidence of how the permutation testing allowed by MDRGPU enables the discovery of combinations of SNPs that are significantly associated with a disease.

4 DISCUSSION

While SNP rs4363506 has been reported as associated with disease in Schymick *et al.* (2007), it did not have a statistically significant effect in Cronin *et al.* (2008) when considered alone (χ^2, $P = 0.18$) and would have failed to replicate without considering pairwise effects. SNP rs6014848 has not previously been described as associated with sporadic ALS, although it shows main effects (uncorrected χ^2, $P < 0.05$) in both datasets. Greene *et al.* (2009) have shown that SNPs can fail to replicate a significant association when the joint effect of those SNPs is ignored. This is particularly likely when the populations from which patients are ascertained differs. Schymick *et al.* (2007) collected individuals from the USA, while Cronin *et al.* (2008) collected individuals from Ireland. By considering the joint effect of SNPs, MDRGPU discovers a novel association which replicates in an independent dataset. GPUs provide a platform for epistasis analysis in genome-wide data where computational requirements far exceed what CPUs can cost-effectively provide. MDRGPU is a software package for this emerging computing platform that enables human geneticists to tackle analyses previously found to be intractable.

ACKNOWLEDGEMENTS

We thank Peter Andrews for code review. The genotyping of samples was provided by the NINDS. The datasets used for the analyses described in this manuscript were obtained from the NINDS Database through dbGaP accession numbers phs000006.v1.p1 and pht000649.v1.p1. Permission to use these datasets was obtained through application to dbGAP by B.T.H.

Funding: NINDS, ALS Assoc, Muscular Dystrophy Assoc (MDA) (to National Institute of Neurological Disorders and Stroke (NINDS) Repository Motor Neuron Disease/ALS Study); MDA (to The Study of Irish ALS); Irish Institute of Clinical Neurosciences (to The Study of Irish ALS); National Institutes of Health (NIH); (Intramural funding to to The Study of Irish ALS). NIH (grants LM009012, LM010098, AI59694, HD047447 and ES007373).

Conflict of Interest: none declared.

REFERENCES

Barrett,J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Cronin,S. *et al.* (2008) A genome-wide association study of sporadic ALS in a homogenous irish population. *Hum. Mol. Genet.*, **17**, 768–774.

Greene,C.S. *et al.* (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE*, **4**, e5639.

Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.

Ritchie,M.D. *et al.* (2001) Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Schymick,J.C. *et al.* (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **6**, 322–328.

Shriner,D. *et al.* (2007) Problems with genome-wide association studies. *Science*, **316**, 1840–1841.

Sinnott-Armstrong,N. *et al.* (2009) Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Res. Notes*, **2**, 149.

Wang,W.Y.S. *et al.* (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.

Williams,S.M. *et al.* (2007) Problems with genome-wide association studies. *Science*, **316**, 1841–1842.