Cell Research (2010) 20:854-857.

www.nature.com/cr

LETTER TO THE EDITOR

SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder

Cell Research (2010) **20**:854-857. doi:10.1038/cr.2010.68; published online 25 May 2010

Dear Editor,

We developed a GPU-based analytical method, named as SHEsisEpi, which purely focuses on risk epistasis in a genome-wide association study (GWAS) of complex traits, excluding the contamination of marginal effects caused by single-locus association. We analyzed the Wellcome Trust Case Control Consortium's (WTCCC) GWAS data of bipolar disorder (BPD) with 500K SNPs. Our algorithm only used 27 h to finish the exhaustive scan and was more than 300 times faster than the CPU-based analysis on our system. Furthermore, by genotyping the top finding that met our criteria ($P = 5.37 \times 10^{-12}$) and its nearby SNP pairs in another independent 475 BPD patients and 480 normal controls from Chinese Han population, we validated these findings, related with two gene pairs, conferring risk ($\alpha = 0.05$) to BPD. Binary files and source codes of our program can be downloaded at http://analysis.bio-x.cn from the main menu of SHEsis.

GWAS, which is a simplistic, exhaustive, whole genome 'one SNP at a time' analysis approach focused on single-locus marginal effects [1], has led to the discovery of a lot of risk loci of complex diseases. But in most cases, there are more than one pathogenic loci. Much empirical evidence has shown that complex traits are broadly affected by interactions among different loci [2-5]. Theoretically there are a variety of purely epistatic disease models, which display no main effects at all [6]. In this occasion, genome-wide gene-gene interaction analysis is requested to identify hidden multi-locus susceptibility, especially risk interaction effects without contamination of single-locus association.

In this study, we used CUDA (Compute Unified Device Architecture) to develop a GPU-based algorithm, named as SHEsisEpi, which can efficiently scan all pairwise SNP-SNP interactions in GWAS, and report purely risk epistasis without single-locus association contamination. Our algorithm focused on the different interaction of genotype between cases and controls. For a SNP pair, there were nine genotype combinations normally. We built 2×2 contingency tables by each genotype combination and the clump of other combinations, and then calculated the odds ratio. The final statistic EOR was the quotient of such odds ratios in cases and controls. If the given EOR did not significantly deviate from 1, there was no significant difference of interaction of this genotype combination between cases and controls; otherwise, risk epistasis exited. A *Z*-test was used to estimate the *P*-value of EOR. We have shown the mathematical details in Supplementary information, Data S1. We also demonstrated in Supplementary information, Data S2 that our method purely detected risk epistasis, and was not affected by marginal effects of single-locus association. More details on epistasis are discussed in Supplementary information, Data S3.

We implemented our algorithm on a CUDA platform. As only NVIDIA's GPU higher than GeForce8800 series supports CUDA, our platform is composed of two GTX285 video cards with 1G graphic memory individually, and an Intel i7-920 CPU with four cores (2.67 GHz) along with 12G DDR3 1600 memory, and the operating system is Windows Vista 64bit version. The multithread technique was used in order to take advantage of both two GPUs and cover I/O delay. When running the comparative CPU program, another two threads were built and simulated two GPUs' calculation. The essential drivers and libraries could be downloaded from http://www.nvidia.com. We provided more information about GPU programming in Supplementary information, Data S4. Figure 1 shows the flow diagram for the whole process to implement our algorithm on GPU.

Genotype data (Affymetrix 500K) of Caucasian from WTCCC [7] were divided into 987 input pairs each with 512 SNPs because it is unpractical to load all SNP data into memory at the same time (e.g. SNPs indexed from

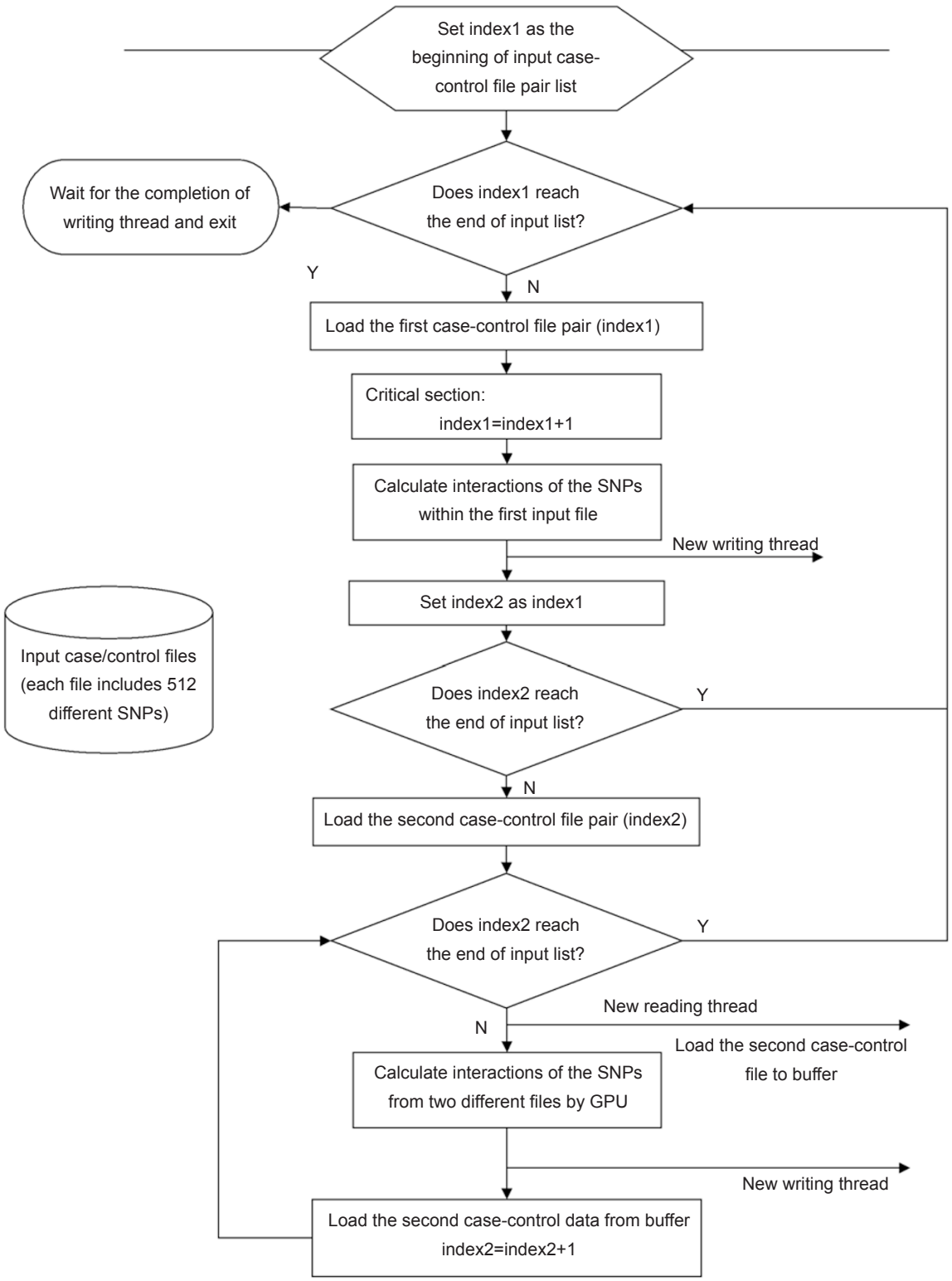

Figure 1 Flow diagram for each of the processing thread. The multithread technique was used to take advantage of both two GPUs. This DFD showed the process of each thread. We used critical section to prevent concurrent access to shared variables. I/O process was also hidden by independent reading/writing threads.

1 to 512 are included in the 1st file, 513 to 1 024 are included in the 2nd file, and so on). We set the number as 512, since it was the maximum number of threads that a block could take under CUDA. In the terminal of a certain chromosome, the quantity may be smaller than 512.

In the accelerator test, eight random file pairs of the whole data set were selected to run on our system. The GPU-based algorithm running on two GPUs was more than 300 times faster than the CPU-based version running on 2 i7-920 CPU cores (More information was provided in Supplementary information, Table S1.). In general cases, considering that there are N pairs of different input files, N internal and $N(N-1)$ external interactions were going to be calculated. With the growth of the input number, time consumption for internal interaction could be ignored and finally the accelerator would be progressive to about 350.

We analyzed WTCCC's BPD GWAS data. It took 27 h to finish the analysis of the whole 500K dataset, while previously a computer cluster composed of 20 CPU cores (Intel Xeon 2.8 GHz) took 32 days. We did not find any global significance ($P < 10^{-13}$ by Bonferroni's correction) during scanning. In order to select candidate SNP pairs for replication study, we ordered the pairs by their P-values and set the following basic criteria considering our limited experimental resources: (1) each SNP in the pair must be on different chromosomes to avoid interaction caused by linkage disequilibrium; (2) each SNP in the significant pair must have a minor allele frequency higher than 0.05 in WTCCC's data; (3) as WTCCC's data included seven different diseases, we excluded SNP pairs that were significant in more than two different diseases, as they might be caused by the deviation in control set.

Based on these criteria, significant SNP pairs were examined one by one, starting from the pair with the lowest P -value. The first pair that fit all criteria was rs10124883-rs178069 ($P = 5.37 \times 10^{-12}$), and their related genes were *ASTN2* and *SNAP29*, respectively. rs6004133, which is located in *PIK4CA*, was close to rs178069 in position, and SNP pair rs10124883-rs6004133 ($P = 6.49 \times 10^{-12}$) also fit all the criteria. Therefore, we decided to further validate these two pairs. In the replication study, we added three SNPs, rs10123629, rs165596 and rs165730, for each gene, to further study the interaction between *ASTN2* and *SNAP29*, and the interaction between *PIK4CA* and *ASTN2*.

A total of 475 unrelated BPD patients (255 males and 220 females, mean age 37.4 ± 9.9 years) as well as 480 normal controls (280 males and 200 females, mean age 36.4 ± 7.2 years) were selected from the Chinese Han population. All the patients met the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition criteria and were ethnic Han Chinese in origin with signed informed consent. The peripheral blood sample was obtained from every subject for DNA extraction, which was performed using QuickGene DNA whole blood kit L (FUJIFILM). Genotyping experiments were carried out by Shanghai Generay Biotech Co., Ltd. (http://www.generay.com.cn/) using allele-specific multiple ligase detection reactions.

The Hardy-Weinberg equilibrium tests were carried out by SHEsis [8, 9], and we provided the results in Supplementary information, Data S5. Both of the significant findings from genome-wide scan, rs10124883-rs178069 ($P = 0.026$) and rs10124883-rs6004133 ($P = 0.021$), were replicated in our independent sample set. We also found rs10124883-rs6004133 ($P = 0.027$), rs10124883-rs165730 ($P = 0.038$), rs10124883-rs165596 ($P = 0.035$), and rs10124883-rs178069 ($P = 0.031$) to be risk epistasis pairs in our independent sample set (Details of the replication study are shown in Supplementary information, Table S2). Our results indicate that interactions between neurological pathway-related genes *ASTN2* and *SNAP29* or *ASTN2* and *PIK4CA*, or other genes in linkage disequilibrium with any genes in these pairs, were risk factors of BPD, in both Caucasian and Chinese Han populations. Further studies are needed to elucidate the etiology of this risk genetic interaction. We have provided more details in Supplementary information, Data S6.

Acknowledgments

This work was supported by grants from the Hi-Tech Research and Development Program of China (2006AA02A407, 2009AA022701), Shanghai Changning Health Bureau program (2008406002), and Shanghai Municipal Health Bureau program (2008095). This study used data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available at http://www.wtccc.org.uk.

Xiaohan Hu[1, *], Qiang Liu[1, *], Zhao Zhang[1, *], Zhiqiang Li[1], Shilin Wang[2], Lin He[1, 3, 4], Yongyong Shi[1]

[1]Bio-X Center and Affiliated Changning Mental Health Center, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai 200030, China; [2]School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200030, China; [3]Institute for Nutritional Sciences, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [4]Institutes of Biomedical Sciences, Fudan University, 138 Yi Xue Yuan Road, Shanghai 200032, China

*These three authors contributed equally to this work.
Correspondence: Yongyong Shi[a], Lin He[b]
[a]E-mail: shiyongyong@gmail.com
[b]E-mail: helinhelin@gmail.com

References

1 Ritchie MD. Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis. *Genome Med* 2009; **1**:65.
2 Mackay TF. Quantitative trait loci in Drosophila. *Nat Rev Genet* 2001; **2**:11-20.
3 Williams SM, Haines JL, Moore JH. *T*he use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? *BioEssays* 2004; **26**:170-179.
4 Segre D, Deluna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. *Nat Genet* 2005; **37**:77-83.
5 Sing CF, Davignon J. Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am J Hum Genet* 1985; **37**:268-285.
6 Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002; **70**:461-471.
7 The Wellcom Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3 000 shared controls. *Nature* 2007; **447**:661-678.
8 Shi YY, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res* 2005; **15**:97-98.
9 Li Z, Zhang Z, He Z, *et al.* Partition-ligation-combination-subdivision EM algorithm for haplotype inference with multiallelic markers: update of the SHEsis (http://analysis.bio-x.cn). *Cell Res* 2009; **19**:519-523.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)