Original Paper

Human Heredity

Hum Hered 2012;73:220–236
DOI: 10.1159/000341885

Received: January 14, 2012
Accepted after revision: July 13, 2012
Published online: September 4, 2012

GLIDE: GPU-Based Linear Regression for Detection of Epistasis

Tony Kam-Thong[a, c] Chloé-Agathe Azencott[a] Lawrence Cayton[b]
Benno Pütz[c] André Altmann[c] Nazanin Karbalai[c] Philipp G. Sämann[d]
Bernhard Schölkopf[b] Bertram Müller-Myhsok[c] Karsten M. Borgwardt[a]

[a]Machine Learning and Computational Biology Research Group, Max Planck Institutes Tübingen, and [b]Department for Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, and [c]Statistical Genetics and [d]Neuroimaging Research Group, Max Planck Institute of Psychiatry, Munich, Germany

Key Words
Epistasis · Genome-wide interaction analysis · Computational biology · Graphics processing units

Abstract
Due to recent advances in genotyping technologies, mapping phenotypes to single loci in the genome has become a standard technique in statistical genetics. However, one-locus mapping fails to explain much of the phenotypic variance in complex traits. Here, we present GLIDE, which maps phenotypes to pairs of genetic loci and systematically searches for the epistatic interactions expected to reveal part of this missing heritability. GLIDE makes use of the computational power of consumer-grade graphics cards to detect such interactions via linear regression. This enabled us to conduct a systematic two-locus mapping study on seven disease data sets from the Wellcome Trust Case Control Consortium and on in-house hippocampal volume data in 6 h per data set, while current single CPU-based approaches require more than a year's time to complete the same task.

Tony Kam-Thong and Chloé-Agathe Azencott are equal first authors. Bertram Müller-Myhsok and Karsten M. Borgwardt are equal last authors.

Introduction

Current genotyping and automated phenotyping technologies generate massive amounts of biological data, making it possible to study the relationship between genotype and phenotype at an unprecedented level of detail. While genome-wide studies that associate single DNA locations with phenotypic outcomes have become standard, they unfortunately often fail to explain much of the phenotypic variation in complex diseases [1]. It is widely accepted [1] that one step towards unveiling the missing heritability is to consider interactive effects, a phenomenon called epistasis, across the whole genome. Examples of phenotypes for which synergistic effects between gene loci have indeed proven a reliable predictor variable of the phenotypic outcome include diseases such as type 1 and type 2 diabetes [2, 3], inflammatory bowel disease [4] and hypertension [5]. More recently, genetic interactions have been studied and observed in the context of cancer cell proliferation. Several examples detailing the different nature of genetic interactions enhancing or suppressing cancer mutations are listed in Ashworth et al. [6], and new therapeutic treatments have been proposed to target these interactions. In addition, epistatic effects have also been observed in intermediate phenotypes gained by neuroimaging such as working memory-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2012 S. Karger AG, Basel
0001–5652/12/0734–0220$38.00/0

Accessible online at:
www.karger.com/hhe

Bertram Müller-Myhsok
Statistical Genetics Group, Max Planck Institute of Psychiatry
Kraepelinstrasse 2
DE–80804 Munich (Germany)
Tel. +49 893 062 2246, E-Mail bmm@mpipsykl.mpg.de

related brain activation [7]. It is important to note that the presence of statistical epistasis can help generate new hypotheses but would require an in-depth investigation of the underlying molecular mechanisms involved to substantiate the findings.

Unfortunately, the detection of gene-gene interactions in genome-wide association studies (GWAS) requires a massive amount of computation. In human GWAS, the number of single-nucleotide polymorphisms (SNPs), that is, single DNA base pairs where genetic variation is observed, is typically in the order of 10^6–10^7. Hence, the number of pairs of SNPs that need to be examined can be in the order of 10^{12}–10^{14}, and data are likely to be collected on as many as hundreds of thousands of individuals in the largest study cohorts. The software and computational strategies employed for data analysis need to be redesigned to accommodate for such large amounts of information.

A possible approach to alleviating the computational burden of epistasis detection consists of reducing the study to a subset of the available SNPs. One can, for example, limit the search space to pairs of SNPs previously singled out by single-locus association studies [8]. This, however, fails to capture high-significance pairs with low main effects. It is also possible to focus on SNPs belonging to relevant pathways, but this does not lead to the discovery of previously unknown biological knowledge. Other space-pruning methods have been recently developed but are limited to either homozygous SNPs [9] or discrete phenotypes and genotypes [10].

Another strategy is to use technological advances and, in particular, to leverage the power of the multiple cores available on graphics processing units (GPUs) to dramatically speed up exhaustive searches. Several software tools designed to perform epistasis searches on GPUs, such as SHEsisEpi [11], EPIBLASTER [12], EPIGPUHSIC [13] and GBOOST [14], have recently been proposed and demonstrated substantial advantages of the use of GPU in this application. However, they are either restricted to binary or discrete phenotypes, which limits the scope of data sets they can analyze, or neglect main effects, which hinders the overall interpretation of their results.

More recently, Hemani et al. [15] developed a GPU-based exhaustive search method that can be applied to quantitative phenotypes. Despite its alluring performance in speed, the application of the proposed F test is ultimately limited to SNP-pair combinations with a 3 × 3 contingency table of possible genotype combinations. In other words, information of the SNP-pair combination can only fall into 9 possible classes. As a result, this method is not applicable to real-number input values, such as in imputed genotypes. Therefore, our proposed method aims to be general enough to be applicable to pairwise epistasis studies of various real or continuous value predictor inputs (genetic and environmental factors) related to the phenotypic output. Furthermore, there is a need for the development of a new GPU-based software tool which not only achieves rapid computational speed but also generates a set of results that all researchers in this field can easily interpret.

Here, we present GLIDE, a high-performance GPU-based implementation of a systematic epistasis search. It computes the statistical scores of the intercept, univariate and interaction coefficients of a linear regression, addressing all the limitations of previous GPU-based methods. The implementation runs on standard GPUs, yielding an inexpensive and exceptionally fast method for epistasis detection. While the actual computation time depends on the technical specifications of the computer used, GLIDE is consistently hundreds to thousands of times faster than state-of-the-art implementations on central processing units (CPUs). In particular, GLIDE makes it possible to conduct a systematic epistasis detection study on the GWAS data published by the Wellcome Trust Consortium [16] in about 6 h per data set using a relatively inexpensive setup of 12 GPUs. In a single-core CPU-based setup, a similar approach would take roughly 1 year to complete. Although an alternative approach, BOOST [17], can be utilized to speed up the process on standard CPUs, it requires that phenotypic and genotypic information is in the discrete domain. By contrast, GLIDE's ability to analyze imputed genotypic data and quantitative phenotypes enables us to search for genotypic interactions predicting hippocampal volume in humans, using a data set of 567 subjects and over a million SNPs collected at the Max Planck Institute of Psychiatry.

Materials and Methods

Linear Regression

Let X be an $m \times n$ matrix, consisting of m subjects and n SNPs. Let y be an $m \times 1$ phenotype vector. We wish to discover correlations between SNP pairs and the phenotype. For each SNP pair $(i, j) \in \{1, ..., n\}$, define the $m \times 4$ matrix

$$X^{ij} = \begin{bmatrix} | & | & | & | \\ 1 & x_i & x_j & x_i \circ x_j \\ | & | & | & | \end{bmatrix}, \qquad (1)$$

where x_i is the i-th column of X (i. e., the i-th SNP over all subjects) and $x_i \circ x_j$ is the element-wise product of the i-th and j-th SNP columns.

For each SNP pair, we wish to estimate a length-four coefficient vector α^{ij} such that $X^{ij}\alpha^{ij} \approx y$.

We solve this estimation problem using the standard linear regression formula $\alpha^{ij} = (X^{ij\top} X^{ij})^{-1} X^{ij\top} y$.

The estimated output phenotype vector $\hat{y}$ based on α^{ij} is then $\hat{y}^{ij} = X^{ij}\alpha^{ij}$ with a residual sum of square error

$$Residual^{ij}_{SSE} = \sum_{k=1}^{m}\left(y_k - \hat{y}^{ij}_k\right)^2. \qquad (2)$$

To determine whether the estimated interaction term is significantly different from zero, we perform a t test with $m - 4$ degrees of freedom. The t score of the interaction coefficient is given by:

$$\frac{\alpha^{ij}_4}{\sqrt{\frac{Re\,sidual^{ij}_{SSE}}{m-4} \times \left[\left(X^{ij\top} X^{ij}\right)^{-1}\right]_{4,4}}}. \qquad (3)$$

GPU Implementation

GPUs are composed of several hundred lightweight processing units. These devices provide impressive computational power but are only effective for tasks that decompose into many subproblems that can be solved in parallel. Fortunately, our problem is composed of many independent regression tasks, hence fits naturally onto the GPU architecture.

In the GPU programming model, each processing unit executes a thread. These threads are grouped in blocks; within each block, threads can cooperate through execution synchronization and efficient low-latency memory sharing. Leveraging this block structure to reduce memory accesses is crucial for performance.

GLIDE associates each thread with a single regression problem. These threads are then collected into blocks such that threads within a block can share access to a subset of the SNP-subject matrix X. In particular, each two-dimensional block of size $BS \times BS$ loads $2 \times BS$ columns of X and solves all pairwise linear regression problems on the corresponding $BS \times BS$ pairs of SNPs.

GLIDE is written in the C programming language using NVIDIA's CUDA extension. Further details about the GPU implementation can be found in Appendix A.

Hardware and Software Setup

We conducted this study using 12 NVIDIA GTX 580 GPUs. These cards have 16 streaming multiprocessors, each holding 32 processors, yielding a total of 512 GPU cores. They support double-precision floating-point calculations. The host machine is running on an Intel Core i7 920 with a 2.66-GHz CPU host using 12 GB of DDR3 RAM. GLIDE is compiled using the NVIDIA NVCC compiler along with GCC version 4.3.1.

Results

Performance Analysis on Synthetic Data

We validate GLIDE and evaluate its performance on data simulated using a normally distributed output phenotype (mean 0 and standard deviation 1) and genotype SNP value in {0, 1, 2} encoding. The SNPs are simulated in Hardy-Weinberg equilibrium (p = 0.05). The number of individuals is set to 10,000 subjects, and 1,000 unique SNPs from two disjoint sets resulting in 1 million unique SNP pairs between the two sets are used. This is a relatively small number of pairs in the GWAS context but it serves to demonstrate the validity of the method.

In order to validate the implementation, we first compare GLIDE to a standard multiple linear regression on the full-rank model (phenotype = $\alpha + \beta x_i + \gamma x_j + \delta x_i x_j$) computed using PLINK [18] (using the *–epistasis flag*). PLINK performs a likelihood ratio test comparing the regression models with and without the interaction term $\delta x_i x_j$. The correlation coefficient between the p values produced by GLIDE and those returned by PLINK is exactly 1, therefore satisfyingly validating the correctness of our implementation.

In order to evaluate the runtime performance of GLIDE over a range of SNP and subject problem sizes, we simulate 1,000 subjects genotyped over 5,008 SNPs. The runtime of GLIDE depends linearly on the number of pairwise interactions to be tested (fig. 1; table 1), for an average speed of about 2.4×10^6 interactions per second on a single GPU.

We then compare the speed of GLIDE with that of the state-of-the-art CPU methods PLINK [18] and FastEpistasis [19]. As the runtime of all three methods scales up linearly with respect to the number of pairwise SNP interactions, we investigate its dependence on the number of subjects in the study (fig. 2; table 2).

The comparison of GLIDE with PLINK reveals the advantage of porting the code onto GPUs. Although the performance depends on technical specifications such as the clock speed, number of cores, cache memory and current load on the system, the speed-up factor of GLIDE over PLINK epistasis consistently remains roughly 2,000 (fig. 2; table 2).

The newly released FastEpistasis method [19] extends the PLINK epistasis module to distribute the work over a multi-CPU environment. Its speed scales up linearly with the number of CPU cores used [19]. The overall speed-up factor of GLIDE over FastEpistasis on a single CPU is approximately 250 (fig. 2; table 2), meaning that one would need to use a cluster of 250 CPUs to compute epistatic interactions with FastEpistasis in the same amount of time as required by GLIDE on a single desktop GPU.

Wellcome Trust Case-Control Consortium

In order to test GLIDE on real data, we carried out an epistatic interaction detection study on seven data sets for

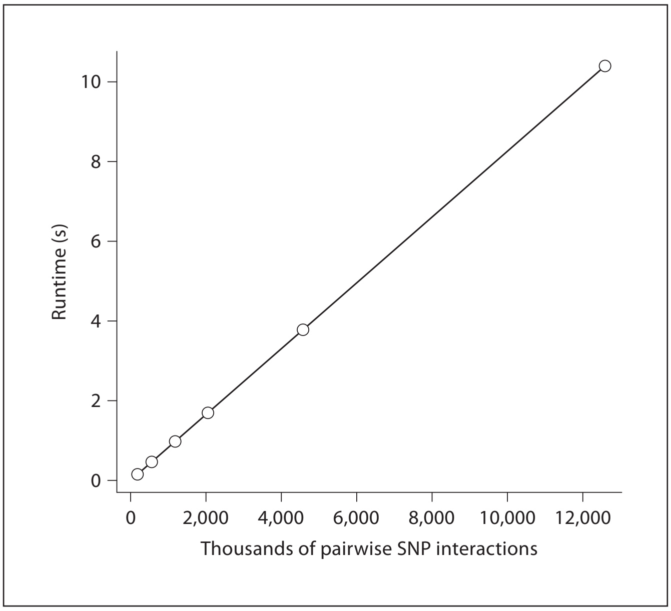


Fig. 1. Runtime of GLIDE, in seconds, as a function of the number of SNP pairs tested.

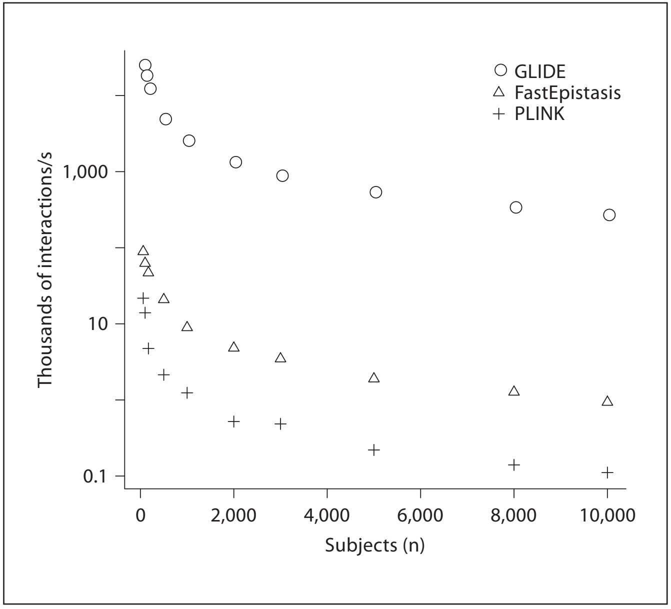


Fig. 2. Speed (thousands of interactions per second) of GLIDE, FastEpistasis (single CPU core), and PLINK as a function of the number of subjects, displayed on a logarithmic scale.

Table 1. GLIDE speed performance against number of pairwise interactions for 1,000 subjects (in number of interactions per second) over 10 runs

Number of interactions	Average speed	SD
130,816	2,396,483	1,653
507,528	2,415,026	599
1,130,256	2,420,631	1,268
1,999,000	2,421,278	477
4,522,528	2,421,397	464
12,537,528	2,421,947	202

common human diseases [bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes] from the Wellcome Trust Case-Control Consortium (WTCCC). Each data set retrieved from the European Genome-phenome Archive (EGA) contains 500,000 SNPs and 5,000 subjects (divided into 2,000 cases and 3,000 controls). We applied the same quality control procedure as described in the original single-locus association study of these sets [16].

Using 0.4 × $n(n - 1)/2$ (where n is the number of SNPs) as correction for multiple hypothesis testing [20] and a significance level of 0.05, we only detected significant interactions for the T1D and the RA data sets.

The original GWAS conducted by the WTCCC [16] has shown strong single-locus associations between both RA and T1D and the human leukocyte antigen (HLA) complex, the most important region of the human genome with respect to infection as well as inflammatory and autoimmune responses [21].

The 170 significant interactions identified by GLIDE for RA and the 3,945 significant interactions detected for T1D all take place between SNPs belonging to this region. While these results may in part be due to the nature of genetic variation in the HLA system, they support the hypothesis that the influence of this system on both clinical conditions is by far not univariate but interactive. For RA, all of those interactions involve SNPs that have significant individual effects and belong to the MHC class II (32.3–33.4 Mb) region, which is consistent with previously reported results [17]. For T1D, 531 of the interactions identified by GLIDE are between SNPs with non-significant individual effects. Most of these interacting pairs reveal interactions between the MHC class I (29.8–31.6 Mb) and MHC class II (32.3–33.4 Mb) regions, which corroborates the hypothesis [22] that genes from both regions should be considered to better understand T1D susceptibility.

Table 2. Compared speeds (in number of interactions per second) of GLIDE, PLINK and FastEpistasis across the number of subjects

Number of subjects	GLIDE speed	PLINK speed	Speed-up factor	FastEpistasis single-core speed	Speed-up factor
60	23,838,755	21,728	1,097	88,950	268
100	17,388,205	13,979	1,244	62,500	278
171	11,683,130	4,750	2,460	46,875	249
500	4,651,763	2,140	2,176	20,833	223
1,000	2,421,486	1,240	1,960	8,929	272
2,000	1,263,967	520	2,419	4,808	263
3,000	839,381	484	1,734	3,472	242
5,000	509,444	220	2,292	1,894	269
8,000	321,630	140	2,285	1,268	254
10,000	255,989	111	2,306	933	274

The performance we observe for FastEpistasis is 2–3 times slower than reported by Schüpbach et al. [19], which is perhaps due to differences in hardware, libraries, and total problem size; regardless, GLIDE is about 100 times faster than the single-CPU FastEpistasis.

Table 3. Physical annotation of the 10 most significant SNP pairs for bipolar disorder in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs17025231	2	101135181	rs4969204	17	74109571	$5.25 \cdot 10^{-11}$
rs8105122	19	15950378	rs17689617	7	40787764	$9.68 \cdot 10^{-11}$
rs4763208	12	10646155	rs17072179	13	48220858	$1.96 \cdot 10^{-10}$
rs6798573	3	21761897	rs4877750	9	83102218	$2.33 \cdot 10^{-10}$
rs4851400	2	101155167	rs4969204	17	74109571	$2.65 \cdot 10^{-10}$
rs1086157	12	105191043	rs2150425	21	41453629	$2.88 \cdot 10^{-10}$
rs1109775	4	102939056	rs4713011	6	26771329	$3.07 \cdot 10^{-10}$
rs1125524	10	7700472	rs1943720	11	84282001	$3.16 \cdot 10^{-10}$
rs1488	6	161508661	rs12814794	12	26331965	$3.12 \cdot 10^{-10}$
rs1125524	10	7700496	rs1943720	11	84282001	$3.33 \cdot 10^{-10}$

Overall, these results match those previously reported by Wan et al. [17] on the WTCCC data sets, where significant interactions between SNPs that are not in linkage disequilibrium nor individually associated with the phenotype were identified only for T1D. The authors additionally report one pair in Crohn's disease, which so far has not been linked to any biological effects. Moreover, this pair was then tested with PLINK and did not pass the corrected significance threshold (Weichuan Yu, pers. commun., June 1, 2011), which is consistent with the fact that it is not picked up by GLIDE in this study.

The physical annotations of the top 10 SNP pair results for the seven data sets are listed in tables 3–9.

Hippocampal Volume

Finally, we conducted an exhaustive search for epistatic interactions associated with hippocampal volume, making full use of GLIDE's ability to handle a quantitative phenotype. The hippocampus is a small but complex bilateral brain structure involved in many cognitive processes, particularly the formation of new memories. An extreme reduction of its volume is a hallmark of Alzheimer's disease, but mild forms of hippocampal volume reductions are also found in patients with schizophrenia or recurrent depression [23]. The hippocampal volume is heritable to some degree, with the heritability estimated from twin studies to be between 40 and 69% [24], and is therefore a good candidate for explicit genetic studies.

Table 4. Physical annotation of the 10 most significant SNP pairs for RA in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs9268877	6	32539125	rs9268645	6	32516505	$2.47 \cdot 10^{-32}$
rs9268877	6	32539125	rs3129872	6	32515131	$1.24 \cdot 10^{-26}$
rs9268877	6	32539125	rs3135342	6	32504593	$1.46 \cdot 10^{-26}$
rs9268877	6	32539125	rs5000563	6	32512113	$2.11 \cdot 10^{-26}$
rs9268877	6	32539125	rs3129877	6	32516575	$4.63 \cdot 10^{-26}$
rs9275134	6	32758590	rs9273363	6	32734250	$6.32 \cdot 10^{-26}$
rs9268877	6	32539125	rs9268831	6	32535726	$2.08 \cdot 10^{-22}$
rs539703	6	32396440	rs3134926	6	32308125	$2.47 \cdot 10^{-21}$
rs3134926	6	32308125	rs4959093	6	32421075	$2.63 \cdot 10^{-21}$
rs574710	6	32396168	rs3134926	6	32308125	$1.31 \cdot 10^{-20}$

Table 5. Physical annotation of the 10 most significant SNP pairs for coronary artery disease in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs11170276	12	51465940	rs273763	18	21451629	$4.07 \cdot 10^{-11}$
rs10876341	12	51454429	rs273763	18	21451629	$1.39 \cdot 10^{-10}$
rs2449393	6	149306948	rs17679635	11	33161304	$1.93 \cdot 10^{-10}$
rs13158763	5	16146630	rs10759217	9	107079850	$3.39 \cdot 10^{-10}$
rs4148202	2	43979470	rs7956731	12	124806808	$3.56 \cdot 10^{-10}$
rs12498185	4	117953156	rs2928579	8	6597571	$3.85 \cdot 10^{-10}$
rs2619283	3	103373585	rs1423977	16	57946190	$4.03 \cdot 10^{-10}$
rs6427623	1	158434421	rs9591697	13	55368450	$4.13 \cdot 10^{-10}$
rs7712927	5	10638111	rs12673840	7	102166072	$4.79 \cdot 10^{-10}$
rs810517	10	80612626	rs11643947	16	5381126	$5.00 \cdot 10^{-10}$

Here, we performed a genome-wide interaction analysis on hippocampal volume automatically determined from high-resolution MRI, with raw volumes corrected for unspecific sources of variance. The bilateral hippocampal volume of 567 subjects was automatically extracted by a method that combines cytoarchitectonic probability maps with optimized segmentation and intersubject coregistration of high-resolution structural MR images, similar to a recent report [25]. Further details on the study sample and phenotype extraction can be found in Appendix B.

Genotypic data were collected using the Illumina 650K chip. We imputed missing genotypic data using MACH [26] on a reference panel Hapmap 3 CEU population sample, as described in detail in Stein et al. [27], and filtered out SNPs with a minor allele frequency (MAF) lower than 5%. We thus obtained a total of 1,075,163 SNPs. The standard single-locus SNP correlation study was carried out first. The top 20 univariate SNP findings are shown in table 10. The lowest p value observed in the univariate GWAS is 2.5×10^{-7}, which is above the critical threshold of the whole genome-wide significance. Of the 20 SNPs with strongest univariate associations, 7 were attributable to the adenosine 3 receptor (ADORA3), a G protein-coupled receptor involved in many intracellular signaling pathways. Of particular interest in the context of hippocampal morphology are animal experiments demonstrating a role of ADORA3 in protecting hippocampal pyramidal cells against hypoxia [28]. Seven additional SNPs were assignable to the potassium-dependent sodium/calcium exchanger SLC24A3, which is highly expressed in thalamic nuclei, hippocampal CA1 neurons, and layer IV of the cerebral cortex, and to SLC6A11, a transporter of the gamma-amino-butyric acid transporter expressed in the hippocampus and involved in brain maturation [29]. All SNPs detected from

Table 6. Physical annotation of the 10 most significant SNP pairs for hypertension in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs16852030	1	159992502	rs10019154	4	114486799	$2.37 \cdot 10^{-11}$
rs9521017	13	108217160	rs17358595	20	15066948	$3.36 \cdot 10^{-11}$
rs9521017	13	108217160	rs17292844	20	15067797	$3.52 \cdot 10^{-11}$
rs10112307	8	135933661	rs12883378	14	28893882	$1.03 \cdot 10^{-10}$
rs9925302	16	85223804	rs1440843	18	25674847	$1.04 \cdot 10^{-10}$
rs731589	8	135929583	rs12883378	14	28893882	$1.35 \cdot 10^{-10}$
rs10024138	4	141121952	rs16905671	10	55615237	$1.35 \cdot 10^{-10}$
rs731589	8	135929583	rs1018542	14	28922805	$1.43 \cdot 10^{-10}$
rs9521036	13	108227221	rs17358595	20	15066948	$1.44 \cdot 10^{-10}$
rs9521036	13	108227221	rs17292844	20	15067797	$1.50 \cdot 10^{-10}$

Table 7. Physical annotation of the 10 most significant SNP pairs for Crohn's disease in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs10809018	9	10025658	rs16907121	9	23331848	$7.70 \cdot 10^{-12}$
rs16907121	9	23331848	rs10809018	9	10025658	$7.70 \cdot 10^{-12}$
rs10887096	10	123916300	rs7318474	13	28027431	$1.02 \cdot 10^{-11}$
rs515309	2	174298192	rs2496731	10	34995393	$6.48 \cdot 10^{-11}$
rs4693426	4	83451117	rs954359	7	152175381	$8.99 \cdot 10^{-11}$
rs6818493	4	83468865	rs954359	7	152175381	$8.99 \cdot 10^{-11}$
rs515309	2	174298192	rs2476995	10	34995323	$9.44 \cdot 10^{-11}$
rs6819282	4	83469210	rs954359	7	152175381	$1.03 \cdot 10^{-10}$
rs4693427	4	83452184	rs954359	7	152175381	$1.06 \cdot 10^{-10}$
rs6554056	4	53562394	rs2609850	22	33175931	$1.55 \cdot 10^{-10}$

the univariate tests were non-hypothesized and have not been directly associated with hippocampal volume in the current literature. Still, the close relationship of the assigned genes with hippocampal physiology provides some first validation by external knowledge.

An exhaustive pairwise test was then performed using GLIDE. The 20 SNP pairs showing the strongest interactions are listed in table 11, and the associated genes (within ±100 kbp) are also reported in table 11. Comparing tables 11 and 10, it can be noted that none of the topmost significant univariate SNPs are involved in the top 20 significant interaction pairs. In other words, all of the most significant pairs would not have been detected if we had first pruned the SNP space based on the univariate tests. In fact, the highest ranked univariate SNP present in table 11 is rs4072698 at a ranking of 54,422 with a univariate p value of 0.052, which is genome-wide insignificant. Furthermore, the p values of the univariate test on the individual loci for the 2 SNPs involved in the top SNP pair, rs10932029 and rs12186557, are also insignificant with 0.23 and 0.89, respectively. This was previously explored by Kam-Thong et al. [12, 13] who showed a poor correlation between the single-locus significance and the two-loci interaction significance. This further stresses the need to adopt an exhaustive search method as opposed to filtering by univariate significance, which will omit significant interactive pairs with low marginal effects.

The two most significant pairs ($p = 2.6 \times 10^{-13}$ and $p = 2.7 \times 10^{-13}$) involve a SNP located in a gene desert of chromosome 5 (rs12186557), paired with either a SNP belonging to the ICOS [inducible (T-cell) costimulator] gene or a SNP located 8 kbp upstream of the CTLA4 (cytotoxic T-lymphocyte-associated protein 4) gene. While the role of rs12186557 remains unknown, both the ICOS and CTLA4 genes are involved in the regulation of the

Table 8. Physical annotation of the 10 most significant SNP pairs for T1D in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs2240063	6	31222724	rs7194	6	32520458	$8.34 \cdot 10^{-25}$
rs3130558	6	31205162	rs2647046	6	32776314	$9.47 \cdot 10^{-25}$
rs3130981	6	31191792	rs2647046	6	32776314	$2.70 \cdot 10^{-24}$
rs2905747	6	31559455	rs9275572	6	32786977	$4.70 \cdot 10^{-24}$
rs3131009	6	31206811	rs2647046	6	32776314	$5.55 \cdot 10^{-24}$
rs3130531	6	31314595	rs6936204	6	32325070	$1.59 \cdot 10^{-21}$
rs3099849	6	31459394	rs9275572	6	32786977	$2.46 \cdot 10^{-21}$
rs3099849	6	31459394	rs2647046	6	32776314	$4.09 \cdot 10^{-21}$
rs2523693	6	31526103	rs9268831	6	32535726	$4.99 \cdot 10^{-20}$
rs2106074	6	31241488	rs9275572	6	32786977	$6.89 \cdot 10^{-20}$

Table 9. Physical annotation of the 10 most significant SNP pairs for type 2 diabetes in WTCCC data

SNP1	Chr	Position	SNP2	Chr	Position	p value
rs10916293	1	224738665	rs9314349	8	27530121	$2.48 \cdot 10^{-11}$
rs2469354	8	3440295	rs10419469	19	41826577	$4.23 \cdot 10^{-11}$
rs2424475	20	22714828	rs1586789	8	126966068	$4.84 \cdot 10^{-11}$
rs2469354	8	3440295	rs10424565	19	41730256	$1.15 \cdot 10^{-10}$
rs287613	1	224793519	rs9314349	8	27530121	$1.23 \cdot 10^{-10}$
rs10792093	11	56910524	rs12157271	2	116666722	$1.57 \cdot 10^{-10}$
rs12134582	1	17138286	rs17651062	6	161931245	$2.18 \cdot 10^{-10}$
rs6549596	3	74590400	rs9711171	2	239513858	$2.45 \cdot 10^{-10}$
rs10896615	11	56920856	rs12157271	2	116666722	$2.48 \cdot 10^{-10}$
rs6961889	7	109923606	rs5771883	22	47367340	$2.54 \cdot 10^{-10}$

adaptive immune system, particularly the development of T-cell functionality and the secretion of different interleukins. Notably, T-cells and inflammatory cytokines have been implicated in neurogenesis and neural plasticity, as demonstrated for hippocampus-dependent tasks in animal models [30, 31].

Genes even more directly linked to brain development appear among the other top significant pairs we report. Gene ZEB2 is particularly interesting as it encodes the zinc finger E-box-binding homeobox 2 protein, which interacts with SMADs, small intracellular signal integrators that act in the transforming growth factor beta signaling pathway. This pathway plays a role for embryonic development in terms of cell differentiation, cell growth, and apoptosis [32]. Furthermore, it is a key modulator of the Wnt pathway, which regulates hippocampal development [33], and a candidate gene study using voxel-based morphometry has associated SNPs in ZEB2 with right temporolateral and hippocampal cortex volume [34]; in addition, temporal lobe abnormalities have been reported in ZEB2 mutations [35]. Further significant interactions were detected for pairs in KIAA1804 [also referred to as mixed linkage kinase 4 (MLK4)] and ZPLD1 (zona pellucida-like domain containing 1). ZPLD1 has been linked to the occurrence of cerebral cavernous malformations [36], but there is limited knowledge on the functionality of MLK4. In addition, three pairs involve TRPM6, a gene encoding a cation channel and expressed in the brain [37]. Two other pairs involve a SNP in protocadherin 8 (PCDH8), which plays a role in cell adhesion in a way specific to the central nervous system [38].

Furthermore, standard statistical genetics analyses were performed in view of substantiating these findings. First, taking a closer look on these top pairs reveals that in fact several pairs have a common SNP member, while

Table 10. The 20 most significant univariate SNPs for hippocampal volumetry

SNP	Position chr:kbp	Gene	Distance kbp	p value
rs4838917	1:111915	ADORA3	–7	$2.05 \cdot 10^{-7}$
rs2364815	1:111907	ADORA3	0	$2.06 \cdot 10^{-7}$
rs17663802	1:111910	ADORA3	–2	$5.03 \cdot 10^{-7}$
rs10776733	1:111909	ADORA3	–1	$6.00 \cdot 10^{-7}$
rs10857896	1:111915	ADORA3	–7	$1.20 \cdot 10^{-6}$
rs1905755	3:10810	SLC6A11	–22	$2.05 \cdot 10^{-6}$
rs1905752	3:10808	SLC6A11	–24	$2.05 \cdot 10^{-6}$
rs10857898	1:111922	ADORA3	–14	$2.15 \cdot 10^{-6}$
rs12566794	1:111923	ADORA3	–15	$2.25 \cdot 10^{-6}$
rs6035224	20:19043	SLC24A3	–97	$2.41 \cdot 10^{-6}$
rs1033814	20:19049	SLC24A3	–91	$4.62 \cdot 10^{-6}$
rs6045840	20:19045	SLC24A3	–95	$5.60 \cdot 10^{-6}$
rs6045843	20:19046	SLC24A3	–94	$5.68 \cdot 10^{-6}$
rs12619086	2:15098			$5.71 \cdot 10^{-6}$
rs2208796	20:19048	SLC24A3	–92	$5.71 \cdot 10^{-6}$
rs4668855	2:15100			$5.72 \cdot 10^{-6}$
rs6431665	2:15100			$5.72 \cdot 10^{-6}$
rs6711699	2:15096			$5.72 \cdot 10^{-6}$
rs12619056	2:15098			$5.73 \cdot 10^{-6}$
rs4668858	2:15100			$5.75 \cdot 10^{-6}$

Coordinates are reported using the hg18 genome map. In the distance column, '–' indicates upstream of the gene and a distance of 0 means in the gene. No gene name means no gene was found within 100 kbp of the SNP.

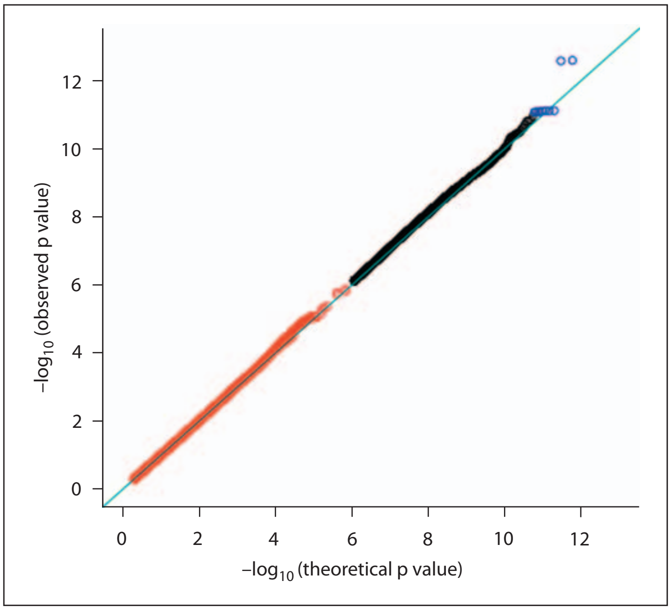
the complementary SNPs are in close physical proximity and most likely in linkage equilibrium. Such is the case for the top two SNP pairs: rs12186557 is the common SNP, and rs10932029/rs11571300 are in close proximity. This is, in fact, catching the same effect. In addition, two pairs can involve 4 unique SNPs, but when both SNPs are in close proximity to the SNPs in another pair, this is also a redundant observation. Such is the case for SNP pairs rs12614080/rs17060595 and rs13392477/rs1333343. As a result, if one filters out redundant observations from the top 20 pairs, only 8 pairs are truly unique. Constructing a model with all 8 unique top pairs (italic type in table 11) to analyze their effects jointly reveals that the p values of the individual SNP pairs are robust, and the fraction of variance that is explained by this joint model containing these 8 pairs is approximately 0.40.

Constructing a Q-Q plot across all test scores to search is practically infeasible for the total amount of SNP pairs involved. Instead, a Q-Q plot of stratified sampling of the top 510,220 (or top 8.8 × 10^{-5}%) observed $-\log_{10}$ p value pairs, and a subset of 500,000 randomly chosen pairs from the remaining lower ranked pairs (or bottom 99.999912%) against a standard normal distribution is constructed. Figure 3 shows no deviation from the unit slope nor unbiased amount of clustering in the upper quantiles, thus indicating that the significant findings are not due to any skewness of the distribution from normality and are as expected.

Finally, it is important to develop a three-dimensional map to visually illustrate the effects of the genotypic interactions with respect to the phenotype. As real valued variables are used in both the genotype and phenotype, a tailored plot is constructed. To this end, a plot containing the raw data points is shown in figure 4 for the very top SNP pair finding. A density plot of the data points is also shown on the SNP-SNP plane. In addition, to investigate the improvement on the quality of the fit due to the interaction term, the fitted plane based on just the linear combination of the two SNP models is plotted as a blue grid (color refer to online version only). Moreover, the fitted surface taking the interaction term into account is also shown in yellow. Comparing the scatterness of the raw data points to the plane reveals that the overall behavior and many of the local undulations of the data points cannot be captured by a simple flat surface. A resulting fitted surface from the model embedding the interaction term proves to be a better fit to the behavior of the data points, as illustrated. In addition, by rounding the imputed genotype data to the nearest integer, 3 × 3 tables for the number of subjects and phenotypic means in each cell are shown in tables 12 and 13, respectively. It can be shown that the result obtained is in part driven by a small number of outliers at the extreme points.

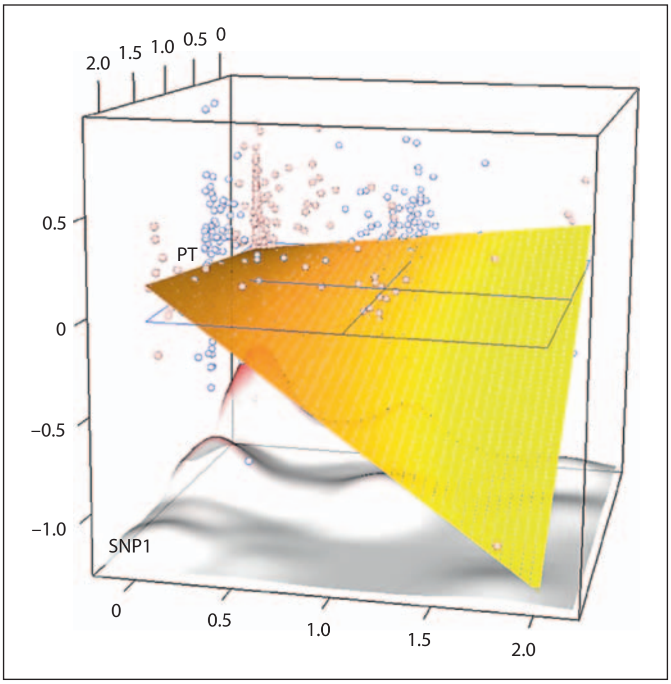
Discussion

One step towards revealing the missing heritability in complex traits is to search for epistatic effects and to map phenotypic variation to pairs of genetic loci. We implemented a fast two-locus genome-wide interaction detection algorithm, which performs an exhaustive SNP-SNP interaction search on typically sized large-scale studies in 6 h on relatively inexpensive GPUs. Unlike other available GPU-based search methods [11–15], GLIDE can be applied to quantitative phenotypes and real-valued genotypes. The ability to work with real-valued numbers opens new opportunities to analyze interactions with other sources of predictor variables such as environmental factors. In large cohort studies where data collection



Color version available online

Fig. 3. Q-Q plot of stratified sampling of the hippocampal volume study. Top 510,220 (or 8.8 × 10^{-5}%) observed –$\log_{10}$ p value pairs in black, subset of points from the remaining lower ranked pairs in red, unit slope line in cyan, and top 20 pairs presented in table 8 in blue (colors refer to online version only).



Color version available online

Fig. 4. Genotype-phenotype map for the top pair in the hippocampal volume study. Raw data points (red/blue spheres) with corresponding density plot (bottom), fitted plane from additive univariate model (blue grid), and fitted surface from the interaction model (yellow) (colors refer to online version only).

Table 11. Physical annotation of the 20 most significant SNP pairs for hippocampal volume

SNP	Position chr:kbp	Gene	Distance kbp	p value
rs10932029 *rs12186557*	2:204510 5:104962	ICOS	0	2.60 • 10^{-13}
rs11571300 rs12186557	2:204455 5:104962	CTLA4	+8	2.69 • 10^{-13}
rs1294230 *rs2063640*	1:231586 3:103686	KIAA1804 ZPLD1	+1 +2	7.77 • 10^{-12}
rs1294229 rs2063640	1:231587 3:103686	KIAA1804 ZPLD1	+2 +2	7.80 • 10^{-12}
rs1294228 rs2063640	1:231586 3:103686	KIAA1804 ZPLD1	+1 +2	7.82 • 10^{-12}
rs1294226 rs2063640	1:231587 3:103686	KIAA1804 ZPLD1	+1 +2	8.03 • 10^{-12}
rs1294205 rs2063640	1:231591 3:103686	KIAA1804 ZPLD1	+5 +2	8.37 • 10^{-12}
rs1294200 rs2063640	1:231591 3:103686	KIAA1804 ZPLD1	+5 +2	8.45 • 10^{-12}
rs1294233 rs2063640	1:231585 3:103686	KIAA1804 ZPLD1	+1 +2	8.50 • 10^{-12}
rs1294198 rs2063640	1:231592 3:103686	KIAA1804 ZPLD1	+6 +2	8.59 • 10^{-12}
rs6746122 *rs4072698*	2:145034 3:103686	ZEB2 ZPLD1	–40 +4	1.13 • 10^{-11}
rs13392477 *rs17060595*	2:136829 9:76673	 TRPM6	 0	1.42 • 10^{-11}
rs12614080 rs17060595	2:136836 9:76673	 TRPM6	 0	1.52 • 10^{-11}
rs4854951 *rs16970848*	3:180919 16:20998	USP13 DNAH3	0 0	1.55 • 10^{-11}
rs4854951 rs16970847	3:180919 16:20997	USP13 DNAH3	0 0	1.56 • 10^{-11}
rs2254788 *rs9568763*	12:94259 13:52278	VEZT PCDH8	+41 +37	1.62 • 10^{-11}
rs13170855 *rs4731513*	5:146655 7:128167	STK32A CALU	0 0	1.97 • 10^{-11}
rs2658679 rs9568763	12:94252 13:52278	VEZT PCDH8	+34 +37	2.03 • 10^{-11}
rs13392477 rs1333343	2:136829 9:76692	 TRPM6	 0	2.50 • 10^{-11}
rs11857420 *rs7205063*	15:94607 16:85539			2.52 • 10^{-11}

Coordinates are reported using the hg18 genome map. In the distance column, '–' indicates upstream of the gene, '+' downstream of the gene, and a distance of 0 means in the gene. No gene name means no gene was found within 100 kbp of the SNP. SNP pairs in italic type indicates non-redundant pairs.

Table 12. Genotype-phenotype map for the top pair in the hippocampal volume study

	rs12186557		
	288	112	6
rs10932029	110	34	3
	12	1	1

Subject distribution across binned 3 × 3 genotype table.

Table 13. Genotype-phenotype map for the top pair in the hippocampal volume study

	rs12186557		
	−0.182	0.066	0.183
rs10932029	0.030	−0.130	−0.578
	0.108	−0.705	−1

Mean hippocampal volume distribution across binned 3 × 3 genotype table.

is done across heterogeneous platforms, the ability to work with imputed genotype data is crucial. In turn, the larger sample size will help improve the power. Furthermore, the proposed method generates easily interpretable test statistics for researchers to work with.

GLIDE implements the same algorithm as PLINK epistasis analysis and is faster by a factor of 2,000. Even the optimized multicore version of PLINK, FastEpistasis, requires roughly 250 CPU cores to run as fast as GLIDE on a single GPU. A GPU can be housed in a basic desktop computer; in contrast, 250 CPU cores require a computer cluster, along with the significant space, power, and management costs that accompany it. Indeed, the costs of the CPU processors alone dwarf the cost of a GPU by an order of magnitude. The significant benefits of the GPU setup render it possible to routinely perform exhaustive two-locus GWAS.

It is important to note that, under certain conditions, the algorithm developed in BOOST [17] using Boolean representation of the genotype data which in turn allows for quick Boolean operations has a performance advantage over our proposed method. The BOOST algorithm has overcome the problem of non-closed form solution for the test of interaction in the logistic regression model by first implementing a screening stage where the Kirkwood superposition [17] was used as an approximation.

At 5,000 subjects, the single-CPU core BOOST solution alone is approximately 1.9 times faster than GLIDE. The extension of this method to GPUs as described in Yung et al. [14] is approximately 75 times faster than GLIDE. However, GLIDE can be substituted or complemented by GBOOST if, and only if, both the phenotype and the genotype data are in the discrete domain. In other words, the phenotype must be of dichotomous nature and the genotype must be non-imputed for GBOOST to be applicable and for its speedup to be exemplified. GLIDE overcomes these limitations by extending the tool to be versatile enough to solve a general linear regression. In addition, GLIDE can be used to solve for linear regression when a dichotomous phenotype is under investigation, as shown in the WTCCC results. This is performed by first transforming the case-control phenotype with the logarithm of the odds ratio between the two classes, the logit transformation. As the genotype is imputed, information for all subjects is present for all SNPs, thus the odds ratios do not vary from one SNP pair to another. Although there are merits in adopting the logistic regression, results obtained from the suggested linear regression approach have been shown in the literature [39, 40] to be asymptotically similar to those obtained with the logistic regression approach when the sample size is large enough for the residuals to be normally distributed. In addition, the maximum likelihood approach corresponds to the least squares approach when the residuals are normally distributed. Thus, as the conditions are met in our study, the measured statistical scores for the interaction coefficient using the linear regression approach with the logit has been adopted for dichotomous phenotype WTCCC data sets.

A crucial feature of GLIDE is that it can directly be applied to continuous phenotypes. As an illustration of this new opportunity, we explored the whole genome for epistatic effects on the hippocampal volume in humans, detecting two promising interactions that involved the ICOS gene (known to play an important role in cell-cell signaling, immune responses, and cell proliferation regulation) or the CTLA4 gene (also involved in the immune response and linked to autoimmune disorders). These findings are striking in the light of accumulating evidence for the impact of immune processes on neuroplasticity [30]. The interacting SNP we identified in both cases, rs12186557, does not, to current knowledge, belong to a known gene; its closest gene, located about 600 kbp upstream, is the pseudogene RAB9BP1. Such result pattern is not unusual in hypothesis-free genetic association studies, with a more recent example being a whole-genome association study on major depressive disorder that revealed polymorphisms not mapping to any annotated gene but with functional relevance [25].

Most notably, the other SNP interactions we detected displaying interactive effects point to genes for which functionality in brain development is highly plausible, such as ZEB2, which is part of the Wnt signaling pathway in development and was already previously detected in a morphology study performed on a subsample of the data set used here [34], and PCDH8, which is known to play a role in cell adhesion, in particular for the central nervous system [38].

These results illustrate how GLIDE can be used to efficiently perform an epistatic search on continuous phenotypes; further analysis must be conducted to explore their biological implications. Following conventional usage, we defined the phenotype as the whole hippocampal volume; however, using the separate cytoarchitectonic subregions could lead to even stronger and more diverse results as it is likely that the development of different subregions depends on distinct genetic processes. In addition, GLIDE could be used to investigate brain regions with higher heritability measures than the hippocampus, such as those determined from twin studies [24].

As exhaustive epistatic search involves testing 10^{12}–10^{14} hypotheses on the same data set, further developments will include addressing multiple hypothesis testing correction. The traditional Bonferroni correction is known to be overly conservative as linkage disequilibrium between nearby markers leads to mutually correlated variables. Modifications of this correction have therefore been proposed [20]. Permutation-based statistical tests [10] are, however, more promising; yet, such approaches are computationally burdensome, and we plan to investigate the design of GPU-based implementations that will speed them up.

Appendix A: GPU Implementation Details

The GLIDE implementation is based on a natural mapping between the required output – all pairwise regression coefficients – and the GPU processing elements, termed thread processors. The exact organization of the computation is crucial to the performance of the implementation. In this appendix, we provide an overview of the implementation details.

Principles of GPU Implementation

The GPU hardware consists of a memory bank and a large pool of grouped processing elements. The CUDA programming interface exposes this organization; the processing elements correspond to threads and the groups correspond to blocks. Threads within the same block share a small amount of very fast on-chip memory, and all threads can access the main memory bank. Access to the main memory bank is slow, relative to the throughput of the processing elements, and hence is often a bottleneck in GPU implementations. To avoid this bottleneck, efficient implementations make use of the on-chip memory to reduce main memory traffic.

Organization of the Computation

Recall the problem: there are n SNPs and the algorithm must compute

$$\frac{n(n-1)}{2}$$

coefficient vectors α^{ij}. The calculation of pairwise epistatic SNP interactions between n SNPs corresponds to filling an upper triangular matrix, as illustrated by the grey area in figure 5.

Each thread is responsible for the calculation of a single coefficient vector α^{ij}. The threads are indexed using the same (i, j) indices as the coefficient vectors. The threads (i, j) must access the i-th and j-th SNPs from the GPU memory. Notice, however, that any two threads sharing an index must access the same SNP. Hence, if the threads are grouped into blocks such that many threads within a block share an index, SNPs loaded from the memory can be shared between threads.

Define the blocks to be of size $BS \times BS$; each block computes a consecutive series of α^{ij}s. For example, the first block computes $\alpha^{11}, ..., \alpha^{BSBS}$. Notice that within this block, each SNP vector is used for the calculation of (at least) BS coefficients, which is the re-use behavior we were after. We describe the block-level computations in more detail in the following subsection.

The implementation creates a grid of

$$\frac{n}{BS} \times \frac{n}{BS}$$

blocks (we assume that n is a multiple of BS for simplicity), which can be executed independently of one another. In many cases, the SNP matrix will be too large to fit into the memory of the GPU. In this case, two chunks of size $m \times n_{\mathrm{GPU}}$ of the SNP matrix X are moved to the GPU, and the coefficients for the $n_{\mathrm{GPU}} \times n_{\mathrm{GPU}}$ corresponding SNP pairs are computed. Next, a different pair of chunks is moved to the GPU and the corresponding coefficients are computed, and so on, until the coefficients for all pairs of SNPs have been computed.

Figure 5 summarizes the principle behind CUDA GPU threads cooperation.

Block-Level Computations

To keep notation simple, we describe the block-level computations for a particular block; other blocks are similar. To evaluate the interactions between BS SNPs from $Set1$ (indexed by $1, \ldots, BS$) and BS SNPs from $Set2$ (indexed by $x_{BS+1, \ldots, 2BS}$), a block computes the matrix $A^{1,2}$ defined as:

$$A^{1,2} = \begin{bmatrix} | & \cdots & | & | & \cdots & | & | & \cdots & | & | & \cdots & | & | & | \\ x_1 & \cdots & x_{BS} & x_{BS+1} & \cdots & x_{2BS} & x_1^2 & \cdots & x_{BS}^2 & x_{BS+1}^2 & \cdots & x_{2BS}^2 & 1 & y \\ | & \cdots & | & | & \cdots & | & | & \cdots & | & | & \cdots & | & | & | \end{bmatrix}. \quad (4)$$

$A^{1,2}$ is of dimension $p \times (4BS + 2)$, where p is the number of subjects chosen to be small enough that all their genotypic information can be read at once. It is stored in the shared memory, which is accessible by all threads within the same block.

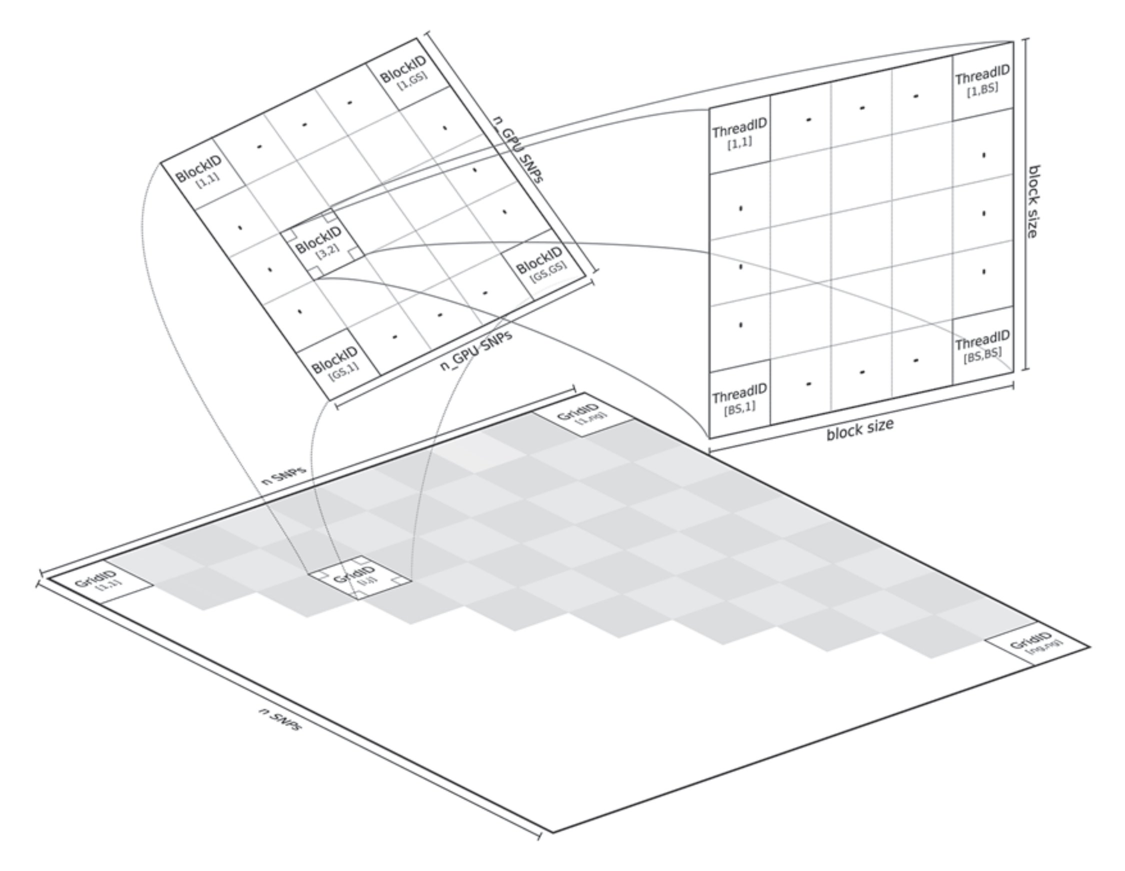


Fig. 5. GPU threads cooperation. The epistatic interactions matrix to be computed, of size $n \times n$, is divided into chunks of size $n_{\mathrm{GPU}} \times n_{\mathrm{GPU}}$. Note that the epistatic interaction matrix is symmetric; it is therefore only necessary to compute the upper triangular values, grayed out on this diagram. The chunks are computed sequentially. Each chunk is divided into blocks of size $BS \times BS$. Each of those blocks is computed in parallel by $BS \times BS$ threads.

Next, the block computes and stores in the shared memory the correlation matrix $T = A^{1,2\top} A^{1,2}$ of dimension $(4BS + 2) \times (4BS + 2)$. This step is implemented as a standard matrix-matrix multiplication on the GPU. T contains all elements necessary to retrieve the correlation matrix $X^{ij\top} X^{ij}$ for every pair of SNPs in $Set1 \times Set2$. Indeed, $X^{ij\top} X^{ij}$ can be written as:

$$X^{ij\top} X^{ij} = \begin{bmatrix} m & x_i \cdot 1 & x_j \cdot 1 & \left(x_i \circ x_j\right) \cdot 1 \\ x_i \cdot 1 & x_i \cdot x_j & x_i \cdot x_j & x_i \cdot \left(x_i \circ x_j\right) \\ x_j \cdot 1 & x_j \cdot x_i & x_j \cdot x_j & x_j \cdot \left(x_i \circ x_j\right) \\ \left(x_i \circ x_j\right) \cdot 1 & \left(x_i \circ x_j\right) \cdot x_i & \left(x_i \circ x_j\right) \cdot x_j & \left(x_j \circ x_j\right) \cdot \left(x_i \circ x_j\right) \end{bmatrix} \quad (5)$$

Computing $X^{ij\top} X^{ij}$ in this manner ensures that all threads are kept busy and maximizes the efficiency of the method.

Once T has been computed, each of the $BS \times BS$ threads i, j finishes the procedure by computing the regression coefficients for one pair of SNPs indexed by $i \in \{1, \ldots, BS\}$, $j \in \{BS + 1, \ldots, 2BS\}$ as well as the corresponding statistical tests. Each matrix $X^{ij\top} X^{ij}$ is of fixed dimension 4×4 and can therefore be inverted analytically. The estimated mean regression coefficients are then tabulated from $(X^{ij\top} X^{ij})^{-1} X^{ij\top} y$. The estimated phenotype $\hat{y}^{ij}$ is then computed, and the variance of the residual σ^2 is estimated by the mean square error:

$$\widehat{\sigma}_{ij}^2 = \frac{\sum_{k=1}^{m}\left(y_k - \widehat{y}_k^{ij}\right)^2}{m-4}.$$

t scores for each estimated coefficient $\alpha_1^{ij}, \ldots, \alpha_4^{ij}$ are computed by dividing the vector of the estimated coefficients α^{ij} by its standard error

$$\sqrt{\widehat{\sigma}_{ij}^2 \times \left(X^{ij\top} X^{ij}\right)_{diag}^{-1}}.$$

The p values are then computed based on the t distribution with (m – 4) degrees of freedom.

The flowchart in figure 6 summarizes the steps taken on both the host machine and the GPU. Furthermore, this flowchart shows the scope operations on a per-grid, per-block, and per-thread basis.

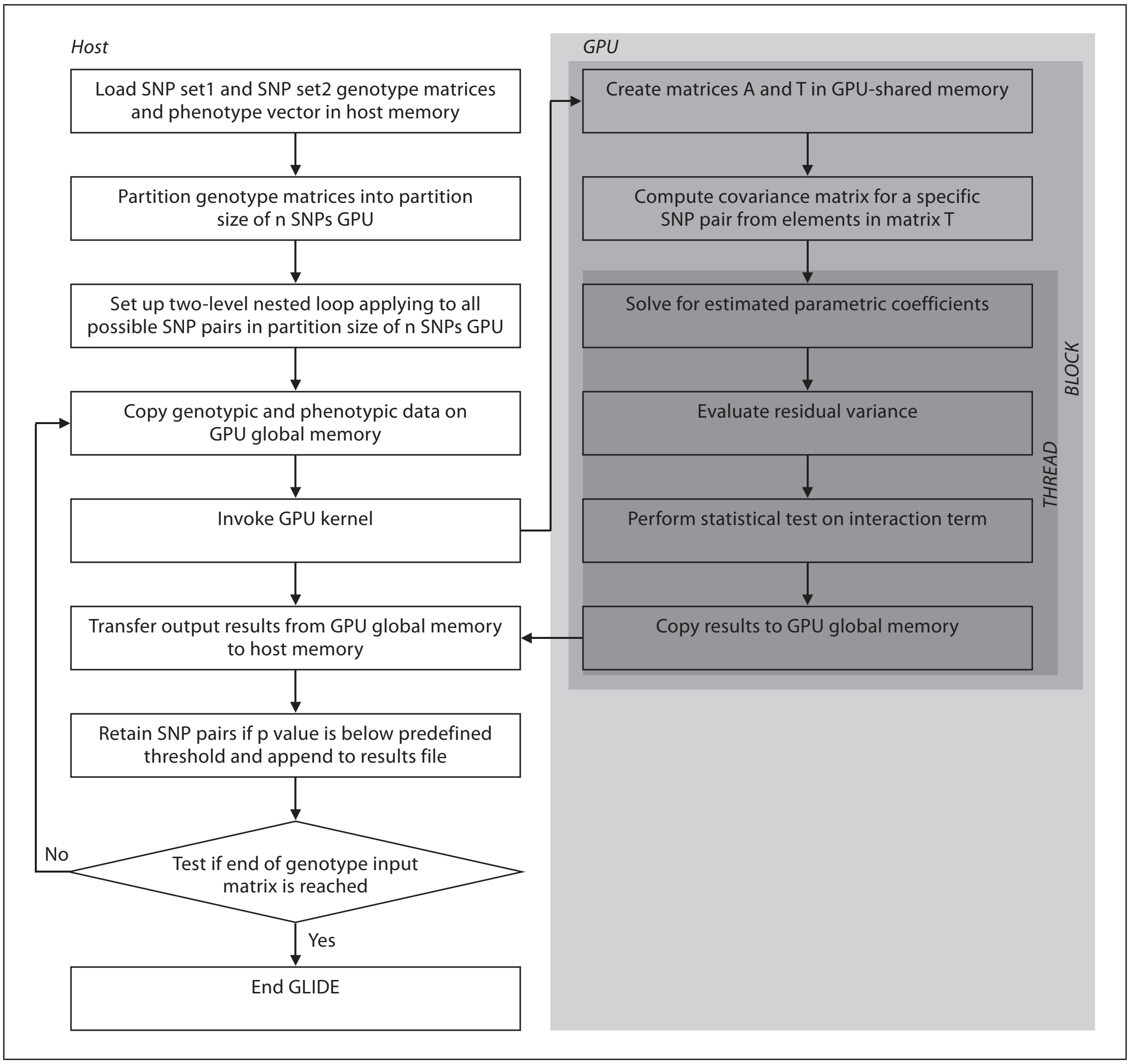

Host
GPU
Load SNP set1 and SNP set2 genotype matrices and phenotype vector in host memory
Partition genotype matrices into partition size of n SNPs GPU
Set up two-level nested loop applying to all possible SNP pairs in partition size of n SNPs GPU
Copy genotypic and phenotypic data on GPU global memory
Invoke GPU kernel
Transfer output results from GPU global memory to host memory
Retain SNP pairs if p value is below predefined threshold and append to results file
Test if end of genotype input matrix is reached
No
Yes
End GLIDE
Create matrices A and T in GPU-shared memory
Compute covariance matrix for a specific SNP pair from elements in matrix T
Solve for estimated parametric coefficients
Evaluate residual variance
Perform statistical test on interaction term
Copy results to GPU global memory
BLOCK
THREAD

Fig. 6. GLIDE's thread cooperation.

Appendix B: MRI Data Acquisition, Sample, and Extraction of Hippocampal Volume

Structural MRI Sample Acquisition of Hippocampal Volume Data and Quality Control

Structural MRI with high-resolution T_1-weighted images adequate for morphometry was available for 204 patients with recurrent unipolar depression and 186 control subjects. MRI was acquired on a 1.5 Tesla clinical scanner (General Electric, Signa Excite, Milwaukee, Wisc., USA) at the Max Planck Institute of Psychiatry in the context of the Munich recurrent unipolar depression and the Munich Antidepressant Response Signature (MARS) studies. The first of these projects gathered data for 200 patients with recurrent depression and 200 control subjects with no history of psychiatric disease, and the second one for more than 170 patients with a depressive disorder. MRI-based criteria for exclusion were signs of territorial brain infarction, gross developmental abnormality, brain neoplasm, and incomplete coverage of the skull, excessive motion artifact, or gross normal variants such as arachnoid cysts that prevent appropriate automated segmentation. The reported sample comprises 567 subjects (379 patients, 188 controls) with a mean age of 48.0 years (SD 13.3) and a gender distribution of men/women of 239/328 (42.1% men). The clinical inclusion and exclusion criteria of both studies have previously been reported in detail [25, 41, 42]. Images used for morphometry were high-resolution T_1-weighted images with optimized grey matter (GM)/white matter (WM)/cerebrospinal fluid (CSF) contrast [sequence details: sagittal T_1-weighted spoiled gradient echo sequence, TR 10.3 s, TE 3.4 ms, 124 slices, matrix size 256 × 256, FOV 23.0 × 23.0 × (14.9–17.4) cm^3, matrix size 256 × 256, voxel size 0.8975 × 0.8975 × (1.2–1.4) mm^3, flip angle 90°; birdcage resonator, post head coil upgrade: TR 9.7 ms, TE 2.1 ms, 124–132 slices, FOV 25 × 25 cm^2, matrix size 256 × 256, voxel size 0.875 × 0.875 × 1.2 mm^3, flip angle 15°].

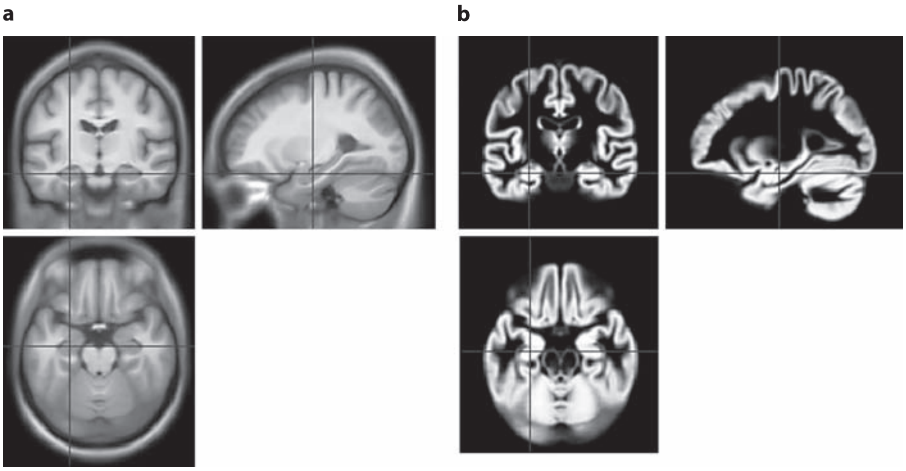


Fig. 7. Coregistration quality of hippocampus volume data. Mean image of all normalized T_1-weighted images (**a**) and GM maps (**b**) as emerging from the DARTEL algorithm.

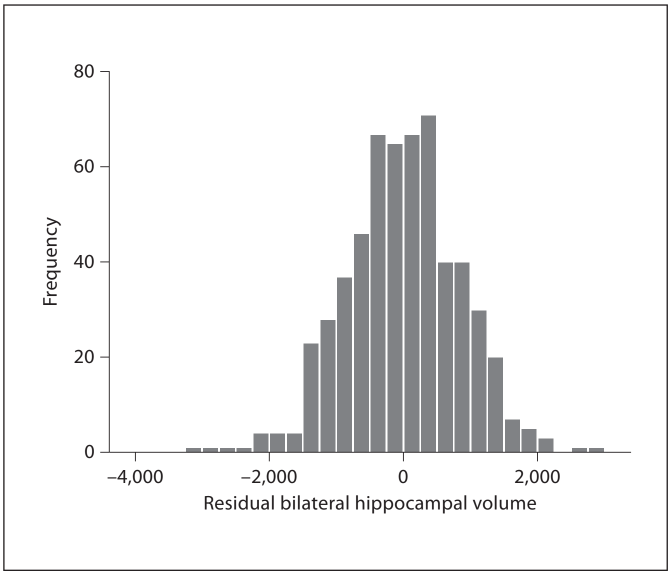


Fig. 8. Distribution of the residualized bilateral hippocampal volumes. The residualized volumes are normally distributed.

Preprocessing of Hippocampal Volume MRI Data

Image preprocessing was performed for voxel-based morphometry to gain GM and WM maps with preserved local volume in stereotactic space. Preprocessing was performed using SPM8, MRIcro graphics, and in-house software written in Matlab 7.0.4 (MathWorks, Natick, Mass., USA) and IDL 6.3. The focus was on optimized coregistration between subjects, which is a prerequisite for valid automated regional volumetry. We employed the diffeomorphic image registration algorithm referred to as DARTEL, which optimizes intersubject alignment [43]. T_1-weighted high-resolution images were subjected to inhomogeneity correction, spatial normalization, and segmentation into GM, WM, and CSF using the unified segmentation algorithm [44], implemented in SPM8, and prior probability maps of the standard SPM8 distribution in MNI152 space (resolution 2 × 2 × 2 mm^3), based on maps from the International Consortium for Brain Mapping. After a first affine alignment of the segmented GM and WM maps to MNI space, these were iteratively coregistered with six generations of GM and WM template pairs in MNI space using linear and non-linear deformations. Default settings were used to specify outer and inner iterations, regularization parameters for each iteration, and optimization settings. Templates at 1.5 × 1.5 × 1.5 mm^3 resolution were based on the DARTEL coregistration of 550 independent healthy adults of the IXI-cohort, available from the VBM8 toolbox. The flow fields resulting from the DARTEL coregistration were applied to segmented native GM, WM, and CSF with Jacobian modulation appended to preserve local volumes. Resulting modulated images were checked visually for consistent alignment and interpolated to a resolution of 1 × 1 × 1 mm^3 to make optimal use of the original resolution of the cytoarchitectonic probability maps.

Automated Regional Volumetry for Hippocampal Volume Data

Based on histologically validated cytoarchitectonic probability maps of hippocampal subregions [45], we derived maximum probability maps (resolution 1 × 1 × 1 mm^3) that comprise the entire left and right hippocampus proper [46] including the cornu ammonis, fascia dentata together with CA4 (referred to as dentate gyrus), and the subiculum. The sum of all modulated GM and WM voxels of the bilateral hippocampal complexes was calculated using an in-house software programmed in IDL. The total intracranial volume was estimated by computing the inverse of the determinant of the affine matrix resulting from an affine coregistration of the brains with a template in MNI space (using the FSL software). This estimate is referred to as eTIV and was used for later residualization. In addition, for proof-of-concept analyses and to detect segmentation failures, total GM, WM, and CSF volumes were extracted and normalized to eTIV. The coregistration quality is documented in figure 7, which shows a mean image of all normalized T_1-weighted images (a) and GM maps (b) resulting from the DARTEL algorithm. Note the sharp delineation of the cortical ribbon in both examples.

Proof-of-Concept Analyses and Residualization

Proof-of-concept analyses confirmed the typical age dependency of normalized total GM, WM, and CSF. In order to gain a single parametric phenotype per subject, the original bilateral hippocampal volumes were residualized against the estimated total intracranial volume, age, squared age, gender, gender × age, gender × squared age, and scanner type in a multiple linear regression model that included all subjects. The residualized volumes are normally distributed (Shapiro-Wilk normality test p = 0.319; fig. 8).

Acknowledgments

The authors thank David Hoehn for insightful and fruitful discussions.

This work was funded in part by the Max Planck Society. Support through the German Ministry for Education and Research (BMBF) through the NGFN (Moods-01GS08145 to B.M.-M.). C.-A.A. is funded by an Alexander-von-Humboldt-Fellowship.

Web Resources

The URLs for data and software presented herein are as follows:

GLIDE is available at http://mlcb.is.tuebingen.mpg.de/Forschung/glide/

SNP annotations are based on Hapmap Genome Browser, http://hapmap.ncbi.nlm.nih.gov/

European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega

SPM8, http://www.fil.ion.ucl.ac.uk/spm/software/spm8

MRIcro graphics software, http://www.sph.sc.edu/comd/rorden/mricro.html

IDL 6.3, http://www.creaso.com

International Consortium for Brain Mapping (ICBM), http://www.loni.ucla.edu/ICBM IXI-cohort, http://www.brain-development.org

VBM8 toolbox, http://dbm.neuro.uni-jena.de/vbm8

FSL software, http://www.fmrib.ox.ac.uk/fsl

References

1 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al: Finding the missing heritability of complex diseases. Nature 2009;461:747–753.

2 Cordell HJ, Wedig GC, Jacobs KB, Elston RC: Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 2000;66:1273–1286.

3 Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A: Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 1999;21:213–215.

4 Cho JH, Nicolae DL, Gold LH, Fields CT, LaBuda MC, Rohal PM, Pickles MR, Qin L, Fu Y, Mann JS, et al: Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. Proc Natl Acad Sci USA 1998;95:7502–7507.

5 Williams SM, Ritchie MD, Phillips JA 3rd, Dawson E, Prince M, Dzhura E, Willis A, Semenya A, Summar M, White BC, et al: Multilocus analysis of hypertension: a hierarchical approach. Hum Hered 2004;57:28–38.

6 Ashworth A, Lord CJ, Reis-Filho JS: Genetic interactions in cancer progression and treatment. Cell 2011;145:30–38.

7 Tan H, Chen Q, Sust S, Buckholtz JW, Meyers JD, Egan MF, Mattay VS, Meyer-Lindenberg A, Weinberger DR, Callicott JH: Epistasis between catechol-O-methyltransferase and type II metabotropic glutamate receptor 3 genes on working memory brain function. Proc Natl Acad Sci USA 2007;104:12536–12541.

8 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–417.

9 Zhang X, Pan F, Xie Y, Zou F, Wang W: COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study. J Comput Biol 2010;17:401–415.

10 Zhang X, Huang S, Zou F, Wang W: TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 2010;26:i217–i227.

11 Hu X, Liu Q, Zhang Z, Li Z, Wang S, He L, Shi Y: SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. Cell Res 2010;20:854–857.

12 Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, Erhardt-Lehmann A, Hemmer B, Rieckmann P, Daake M, Weber F, et al: EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. Eur J Hum Genet 2011;19:465–471.

13 Kam-Thong T, Pütz B, Karbalai N, Müller-Myhsok B, Borgwardt K: Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. Bioinformatics 2011;27:i214–i221.

14 Yung LS, Yang C, Wan X, Yu W: GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 2011;27:1309–1310.

15 Hemani G, Theocharidis A, Wei W, Haley C: EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 2011;27:1462–1465.

16 Wellcome 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.

17 Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, Yu W: BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 2010;87:325–340.

18 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al: PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 2007;41:559–575.

19 Schüpbach T, Xenarios I, Bergmann S, Kapur K: FastEpistasis: a high performance computing solution for quantitative trait epistasis. Bioinformatics 2010;26:1468–1469.

20 Becker T, Herold C, Meesters C, Matthesen M, Baur MP: Significance levels in genome-wide interaction analysis (GWIA). Ann Hum Genet 2011;75:29–35.

21 Shiina T, Hosomichi K, Inoko H, Kulski JK: The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 2009;54:15–39.

22 Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, et al: Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature 2007;450:887–892.

23 Geuze E, Vermetten E, Bremner JD: MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Mol Psychiatry 2005;10:147–159.

24 Peper JS, Brouwer RM, Boomsma DI, Kahn RS, Hulshoff Pol HE: Genetic influences on human brain structure: a review of brain imaging studies in twins. Hum Brain Mapp 2007;28:464–473.

25 Kohli M, Lucae S, Saemann P, Schmidt M, Demirkan A, Hek K, Czamara D, Alexander M, Salyakina D, Ripke S, et al: The neuronal transporter gene SLC6A15 confers risk to major depression. Neuron 2011;70:252–265.

26 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–834.

27 Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann Ø, et al: Identification of common variants associated with human hippocampal and intracranial volumes. Nat Genet 2012;44:552–561.

28 Fedorova IM, Jacobson MA, Basile A, Jacobson KA: Behavioral characterization of mice lacking the A3 adenosine receptor: sensitivity to hypoxic neurodegeneration. Cell Mol Neurobiol 2003;23:431–447.

29 Jursky F, Nelson N: Developmental expression of GABA transporters GAT1 and GAT4 suggests involvement in brain maturation. J Neurochem 1996;67:857–867.

30 Yirmiya R, Goshen I: Immune modulation of learning, memory, neural plasticity and neurogenesis. Brain Behav Immun 2011;25:181–213.

31 Simpson TR, Quezada SA, Allison JP: Regulation of CD4 T cell activation and effector function by inducible costimulator (ICOS). Curr Opin Immunol 2010;22:326–332.

32 Conway SJ, Kaartinen V: TGFβ superfamily signaling in the neural crest lineage. Cell Adh Migr 2011;5:232–236.

33 Miquelajauregui A, Van de Putte T, Polyakov A, Nityanandam A, Boppana S, Seuntjens E, Karabinos A, Higashi Y, Huylebroeck D, Tarabykin V: Smad-interacting protein-1 (Zfhx1b) acts upstream of Wnt signaling in the mouse hippocampus and controls its formation. Proc Natl Acad Sci USA 2007;104: 12919–12924.

34 Inkster B, Nichols TE, Saemann PG, Auer DP, Holsboer F, Muglia P, Matthews PM: Pathway-based approaches to imaging genetics association studies: Wnt signaling, GSK3beta substrates and major depression. Neuroimage 2010;53:908–917.

35 Yoneda M, Fujita T, Yamada Y, Yamada K, Fujii A, Inagaki T, Nakagawa H, Shimada A, Kishikawa M, Nagaya M, et al: Late infantile Hirschsprung disease-mental retardation syndrome with a 3-bp deletion in ZFHX1B. Neurology 2002;59:1637–1640.

36 Gianfrancesco F, Esposito T, Penco S, Maglione V, Liquori CL, Patrosso MC, Zuffardi O, Ciccodicola A, Marchuk DA, Squitieri F: ZPLD1 gene is disrupted in a patient with balanced translocation that exhibits cerebral cavernous malformations. Neuroscience 2008;155:345–349.

37 Fonfria E, Murdock PR, Cusdin FS, Benham CD, Kelsell RE, McNulty S: Tissue distribution profiles of the human TRPM cation channel family. J Recept Signal Transduct Res 2006;26:159–178.

38 Yagi T, Takeichi M: Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. Gene Dev 2000;14: 1169–1180.

39 Hellevik O: Linear versus logistic regression when the dependent variable is a dichotomy. Quality & Quantity 2009;43:59–74.

40 Zhao L, Chen Y, Schaffner DW: Comparison of logistic regression and linear regression in modeling percentage data. Appl Environ Microbiol 2001;67:2129–2135.

41 Hennings JM, Owashi T, Binder EB, Horstmann S, Menke A, Kloiber S, Dose T, Wollweber B, Spieler D, Messer T, et al: Clinical characteristics and treatment outcome in a representative sample of depressed inpatients – findings from the Munich Antidepressant Response Signature (MARS) project. J Psychiatr Res 2009;43:215–229.

42 Binder EB, Salyakina D, Lichtner P, Wochnik GM, Ising M, Putz B, Papiol S, Seaman S, Lucae S, Kohli MA, et al: Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment. Nat Genet 2004;36:1319–1325.

43 Ashburner J: A fast diffeomorphic image registration algorithm. Neuroimage 2007; 38:95–113.

44 Ashburner J, Friston KJ: Unified segmentation. Neuroimage 2005;26:839–851.

45 Amunts K, Kedo O, Kindler M, Pieperhoff P, Mohlberg H, Shah NJ, Habel U, Schneider F, Zilles K: Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat Embryol 2005; 210:343–352.

46 Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K: A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 2005;25:1325–1335.