1 **EPIBLASTER-Fast exhaustive two-locus epistasis detection strategy using graphical**

2 **processing units.**

3

4 **Tony Kam-Thong[1], Darina Czamara[1], Koji Tsuda[2,3,4], Karsten Borgwardt[2,5], Cathryn**

5 **M. Lewis[6,7], Angelika Erhardt-Lehmann[1], Bernhard Hemmer[8], Peter Rieckmann[9],**

6 **Daake M[1], Frank Weber[1], Christiane Wolf[1], Andreas Ziegler[10], Benno Pütz[1], Florian**

7 **Holsboer[1], Bernhard Schölkopf[2], Bertram Müller-Myhsok[1]**

8

9 [1]Max-Planck-Institute of Psychiatry, Kraepelinstrasse 2-10, 80804 Munich, Germany

10 [2]Max-Planck-Institute for Biological Cybernetics, Dept. Schölkopf, Spemannstraße 38

11 72076 Tübingen, Germany

12 [3]Computational Biology Research Center,National Institute of Advanced Industrial Science

13 and Technology (AIST), 2-4-7 Aomi Koto-ku, 1350064 Tokyo, Japan

14 [4]ERATO Minato Project, Japan Science and Technology Agency, 2-12-1-W8-89 O-

15 okayama Meguro-ku, 152-8550 Tokyo, Japan

16 [5]Machine Learning and Computational Biology Research Group, Max-Planck-Institutes,

17 Spemannstraße 38, 72076 Tübingen, Germany

18 [6]King's College London, MRC Social, Genetic and Developmental Psychiatry Centre,

19 Institute of Psychiatry, United Kingdom.

20 [7]King's College London Schools of Medicine, Division of Genetics and Molecular

21 Medicine, United Kingdom.

22 [8]Neurologische Klinik und Poliklinik, der Technischen Universität München Klinikum

23 rechts der Isar, Ismaninger Str. 22, 81675 München, Germany

24 [9]University of British Columbia, Division of Neurology, Dept. of Medicine, 2211

25 Wesbrook Mall, Vancouver BC, V6T 2B5 Canada

26 [10]Institut fuer Medizinische Biometrie und Statistik, Universitaet zu Luebeck,

27 Universitaetsklinikum Schleswig-Holstein, Campus Luebeck, Maria-Goeppert-Str. 1,

28 23562 Luebeck, Germany

29

30

31

32 **Correspondence to:**

33 Bertram Müller-Myhsok

34 Max-Planck-Institute of Psychiatry

35 Kraepelinstrasse 2

36 80804 Munich

37 Germany

38 Tel: +49-(0)89-30622-246

39 Fax: +49-(0)89-30622-642

40 E-mail: bmm[a]mpipsykl.mpg.de

41

42

43 **Running head:** EPIBLASTER fast GPU-based epistasis

44 **Keywords:** Epistasis, Genome-Wide Interaction Analysis, Graphical Processing Unit.

1 **Abstract**

2 Detection of epistatic interaction between loci has been postulated to provide a more in-

3 depth understanding of the complex biological and biochemical pathways underlying

4 human diseases. Studying the interaction between two loci is the natural progression

5 following the traditional and well-established single locus analysis. However, the added

6 costs and time duration required for the computation involved have thus far deterred

7 researchers from pursuing a genome-wide analysis of epistasis. In this paper, we propose a

8 method allowing such analysis to be conducted very rapidly. The method, dubbed

9 EPIBLASTER, is applicable to case-control studies and consists of a two-step process

10 where the difference in Pearson's correlation coefficients is computed between controls and

11 cases across all possible SNP pairs as an indication of significant interaction warranting

12 further analysis. For the subset of interactions deemed potentially significant, a second

13 stage analysis is performed using the likelihood ratio test from the logistic regression to

14 obtain the p-value for the estimated coefficients of the individual effects and the interaction

15 term. The algorithm is implemented using the parallel computational capability of

16 commercially available graphical processing units to greatly reduce the computation time

17 involved. In the current setup and example data-sets (211 cases, 211 controls, 299468

18 SNPs and 601 cases, 825 controls, 291095 SNPs), this coefficient evaluation stage can be

19 completed in roughly one day. Our method allows for exhaustive and rapid detection of

20 significant SNP pair interactions without imposing significant marginal effects of the single

21 loci involved in the pair.

22

1 **Introduction**

2 Understanding the effects of genes on phenotypes and diseases has long been suggested to

3 embed a complex form of interaction as a result of inter-inhibitory and -excitatory effects,

4 with any attempt to explain these effects simply as additive effects of the individual genes

5 being an overly simplistic model which ultimately provides an incorrect view of the genetic

6 influence on the phenotype.

7 The study of interaction between polymorphic loci can stem both from a biological and

8 statistical genetics perspective. The first approach establishes a model based on *a priori*

9 knowledge of how the genes function and interact. The latter, being a 'biological blind'

10 approach, helps draw inferences from previously unknown interdependencies between

11 genes. The ultimate objective, similar to all *black-box* studies, is to merge the conclusions

12 drawn from both approaches but since the observations made cannot be measured at a level

13 more finite than the eventual system output, the former approach is more likely to be

14 refined by first having a solid statistical finding as its basis.

15 As our effort primarily focuses on drawing statistical inference on epistatic

16 actions/interactions between genes, a new method is proposed to help improve our

17 capability to search and sift out significant interactions. This paper will discuss the

18 performance of our method in its current implementation. The results applied to a

19 simulated subset of SNPs and to two real genome-wide datasets recorded from panic

20 disorder and multiple sclerosis studies will be presented, followed by a discussion of some

21 properties of the approach.

22

1 **Materials and Methods**

2 **Overview of the two-stage search strategy**

3 The strategy consists of a two-stage approach. First, a filtering stage using the difference

4 of Pearson's correlation coefficients that performs an exhaustive two-locus interaction

5 multiplicative effects[1] search across all possible pairwise SNP combinations is

6 preformed. This is followed by logistic regression analysis on those subset of pairs deemed

7 significant in the previous stage.

8 **Data representation**

9 Each SNP is represented as integer values ranging from 0 to 2 based on the count of a

10 chosen reference nucleotide of the selected SNP for an allele dosage model, or as 0 or 1

11 depending on the genotype for a dominance or recessivity coding. In the current study, the

12 the allele dosage model is applied. An overall matrix is generated to store the information

13 of all SNPs as column vectors and the recorded values for individual subjects along the

14 rows. Column vectors are then analyzed in pairs and the correlation coefficients are

15 tabulated for cases and controls separately. Correlation coefficients are calculated from a 3

16 x 3 ordered genotype matrix, the genotypes being encoded 0,1,2. The difference between

17 the correlation coefficients in cases and controls is then computed and used as an indication

18 of the SNP pair contributing significantly to the classification between cases and controls

19 (please refer to Equation 1).

20   The first stage of analyzing the difference of correlations approach searches for

21 significant interaction terms. The second stage then computes the fit using a full rank

22 logistic model (Equation 2), including the intercept and additive marginal effects, on the

1 subset of loci pairing deemed significant from the first stage, from which a statistical test

2 can be conducted to test for the coefficient of interaction term being significantly different

3 from zero.

4 **Hardware and software setup**

5 The hardware used in the experimental setup consists of two pairs of commercially

6 available NVIDIA GTX295 GPUs running on a Intel Core i7 920 @ 2.67 GHz central

7 processing unit host (CPU) using 12 GB of DDR3 RAM. The software program is

8 implemented in R (version 2.9.2) with the 'gputools' package beta version 0.1-4 installed, in

9 which the function 'gpuCor' permits correlation coefficients to be tabulated for all possible

10 pairwise interactions across the column vectors using the CUDA (Compute Unified Device

11 Architecture) enabled NVIDIA graphic cards. The graphical card uses its parallel

12 computational capability to process independent evaluations faster than conventional CPU

13 based computation. Since the correlation coefficients between each SNP pair can be

14 tabulated independently, this can take full advantage of the inherent parallel computation

15 done on graphical cards.   The overall time performance depends on the sample size and

16 desired marker coverage. A total evaluation of (number of SNPs choose 2) interactions is

17 typically accomplished within 24 hours for the entire data set (2000 individuals consisting

18 of 1000 cases, 1000 controls with 500000 SNPs) with the available GPU resources and the

19 given results retention criteria. Limitations on the speed can originate from local main

20 memory storage, memory transfer speed and number of on-board GPU cores present.

21 Some data partitioning to take advantage of all current GPU resource is thus required to

22 render this method most efficient. The dataset for the study is first partitioned into blocks

containing 2000 SNPs each that can be handled by the memory on the graphic card.

Hence, for a genome-wide dataset of 500k SNPs, 250 partitions are required.

The process goes through the entire dataset and calculates the correlation coefficients in blocks of 2000 SNPs. The very first correlation analysis in performed is on the first partition to itself, a 'partition-based autocorrelation', resulting in 1999000 unique correlations. The process then increments the partition index of the second partition by one and completes a correlation between two distinct sets of 2000 SNPs, a 'partition-based cross-correlation', to yield 4 million unique results. This process of incrementing the nested loop index is repeated until it reaches the last partition set at which point the top level loop index gets incremented by one. The process can be summarized in the following steps:

1. Partition the data set into size of 2000 SNPs. Note that this number may increase or decrease depending on the number of individuals studied.

2. Set up a two-level nested loop to apply the partition-based correlation for all possible SNP pairs for cases and controls separately.

3. Compute difference of correlation coefficients between cases and controls after each partition-based autocorrelation or cross-correlation is complete.

4. Compute the p-values of each difference given that the distribution of the differences follows a Gaussian.(refer to Results section).

5. Retain only SNP pairs that show a p-value below a selected threshold.

6. Repeat steps 3-5 across all partition pairs.

7. Proceed to stage 2 by performing a logistic regression on the selected pairs.

1   **Results**

2   **Simulated Data**

3   A simulated dataset is generated consisting of 2000 SNPs and a subject size of 5000

4   controls and 5000 cases. This simulated dataset is created without any specific model

5   allowing for any a priori knowledge of which particular pair will be significant, the purpose

6   is demonstrate the validity in the approximation of the resulting logistic regression

7   interaction term p-value to the approximation based on the difference in correlation

8   coefficients. The distribution of the differences of correlation coefficients is noted to

9   exhibit a Gaussian distribution within each partition set, referring to the histogram plot in

10   Figure1. This observation has been examined in greater detail by Gretton et al.[2] stating

11   that when samples are indeed drawn from two different distributions, the distribution of the

12   discrepancy of the chosen function, difference of estimated mean correlation coefficients

13   in this study, will converge to a Gaussian distribution. An additional proof for the

14   difference of correlation coefficients to exhibit a Gaussian distribution can be found in

15   Wellek and Ziegler[3], who have also shown that the variance of any single one difference

16   under the null hypothesis and thus also of the distribution of the sum of all differences is

17   the sum of the reciprocals of the number of cases and controls. For this Gaussianism it is

18   not needed that there are equal numbers of cases and controls.

19   In practice, to test for the significance of each pair, a Z-score is tabulated for each

20   difference within the partition set. This Z-score is computed based on the mean and the

21   standard deviation of all the differences noted within the partition set which is a close

22   approximation to the overall mean and standard deviation given that the partition size is

1   chosen to be large enough typically resulting in a few million pairs for each partition set.

2   Those interactions exhibiting a high overall Z-score are then taken as an indication that the

3   effect of the interaction term of the two SNPs in question is deemed valuable enough to be

4   passed on to the second stage. This filtered subset is then subjected to a second level of

5   mathematical-intensive evaluation using the likelihood ratio test on the logistic regression

6   model.

7   Referring to Figure 2, the p-values of the interaction product term in a general linear fit is

8   plotted against their correlation coefficient differences between cases and controls. To help

9   delineate any logarithmic trend, the p-values are shown as the negative logarithmic values.

10 As shown in Figure 2, there is a strong relationship between the two variables, of a

11 parabolic function in the region centered around the origin to a linear relationship in the

12 region of higher values. The region that is of most interest to the study is the higher

13 numerical value region as the p-values are the smallest and the differences are the largest.

14 Since the differences follow closely a Gaussian distribution (Figure 1), a Z-score threshold

15 can be used to estimate the retention rate. The statistic is then estimated using the fact that

16 the Z-score would follow a standard T-distribution with sufficiently large number of

17 degrees of freedom. A plot comparing the p-values obtained between the approximation

18 and the validation step is illustrated in Figure 3 and demonstrates a high R^2 value of 99.9%.

19   To help address the issues of limited physical disk space and of retaining only those

20 interactions that show strong significance, a Z-score of 4.5 was chosen as the cut-off

21 threshold, which corresponds to a probability of 6.8 x10^{-6} retention rate. Thus, for the

22 partition-based autocorrelation generating ~ 2 million (2000choose2) correlation

1   coefficient differences, the mere top 14 interaction pairs are expected to be retained.

2   Overall, we expect the top $\sim 8.5 \times 10^6$ pairs out of a possible $\sim 1.25 \times 10^{11}$ retained from the

3   first stage in a marker coverage of 500k SNPs.

4   **Real Data**

5   Real genetic data have been recruited from two separate published studies. The first

6   dataset originating from a panic disorder study [4] with a total of 299468 SNPs and 211

7   cases and 222 controls have been retained after standard quality control measures.

8   Computing the difference of correlation coefficients across all pairs and choosing a p-value

9   threshold of 1.0×10^{-5} resulted in a retention of 373153 SNP pairs. Similarly, a second

10 larger dataset from a multiple sclerosis [5] study with a total of 291095 SNPs and 601 cases

11 and 825 controls is also being investigated. Using the same p-value threshold of 1.0×10^{-5},

12 the 407660 most significant SNP pairs are retained upon subjecting it to the first stage.

13   In view of verifying that indeed no significant pairs have been left out in the adopted

14 difference of correlation coefficients stage of our method, a comparison to the p-values of

15 the interaction term in a normal linear regression of all possible SNPs pairs must be made.

16 To perform this brute-force approach in a time efficient manner, we have employed a

17 newly released software tool, FastEpistasis,[6] which is an extension to the PLINK

18 epistasis module capable of distributing the work in parallel on multiple CPU cores, it is

19 important to point out that this method is not working on the difference of Odds Ratio as

20 conducted by the Plink option bearing the same name. The program is meant to be

21 executed on quantitative phenotypes, but the difference in the p-values, which are the

22 relevant measure for this comparison, have been noted to be negligible on several sample

1 SNP pairs (see also Table 1, comparing the FastEpistasis column to the logistic regression

2 interaction term p-value column, but also simulation studies (refer to Supplementary Figure

3 1)). The p-values computed from FastEpistasis is regarded to be the "true" value used for

4 comparison to the approximated method described in stage 1 of EPIBLASTER.

5   Matching the results from SNP pairs with p-values below 1×10^{-6} tested against null from

6 the FastEpistasis with the results obtained from the first stage of EPIBLASTER is

7 performed. From the panic disorder analysis, FastEpistasis produced 37336 SNP pairs of

8 which 36056 of them are also found in the EPIBLASTER stage 1 retained subset (96,5%).

9 The unmatched pairs are indeed examples where EPIBLASTER stage 1 underestimates the

10 p-values and the hard threshold prevents it from being included, thus, these unmatched

11 pairs all are in fact situated around the p-values threshold region and are of lesser

12 significance compared to the others. The plot of the matching pairs are shown on Figure 4

13 and for ease of visualization, it is illustrated as a smoothed color density of the actual

14 scattered points plot. The top 10 most significant pairs from the FatEpistasis approach are

15 listed with greater details in Table 1 along with their annotations in Table 2 and are marked

16 with a dark circle on Figure 4. In order for EPIBLASTER stage 1 to capture all top 10

17 pairs of the "true" approach (FastEpistasis), a p-value threshold of 1.26×10^{-8} must be

18 applied, thus resulting in the top 387 pairs of EPIBLASTER stage 1 to be passed onto stage

19 2. In other words, EPIBLASTER would have produced an additional 377 pairs to be tested

20 in view of capturing the very top 10 true results. In Figure 5, the top 100 SNP pairs of the

21 panic disorder study are marked, which would have resulted in applying a retention

22 threshold for EPIBLASTER stage 1 of 1.67×10^{-7} passing on some 5194 pairs to stage 2.

1   From the multiple sclerosis analysis, FastEpistasis yielded 42731 pairs to have an

2   interaction term with a p-value below 1×10^{-6} of which 42524 pairs (99.5%) are also

3   retained from EPIBLASTER stage 1. The matching pairs along with the respective p-

4   values tabulated using the FastEpistasis method versus the approximated EPIBLASTER

5   stage 1 method are plotted in Figure 6. The top ten pairs are marked in Figure 6 and listed

6   in Table 3 along with the SNP annotations in Table 4. In order for EPIBLASTER to

7   capture the top 10 pairs, it would have required a 48 of its top significant SNP pairs to be

8   carried over to stage 2 where the p-values from logistic regression are tabulated. In

9   addition, in order to capture the top 100 pairs (listed in greater details in Supplementary

10   Table 2), EPIBLASTER would have required the top 19242 pairs obtained from stage 1 to

11   be passed on to stage 2.

12   **Discussion**

13   Although the search is conducted across all possible pairwise SNP interactions, the main

14   interest is to delineate interactions between unlinked loci that influence the illness. In the

15   first stage, the difference of Pearson's correlation coefficients, tabulated from the SNP pair,

16   is taken between controls and cases across all possible interactions. In addition, this step

17   can also incorporate replicating for significant association across two or more independent

18   studies using a number of subjects weighted meta-analysis during the actual run. In the

19   current experimental setup with a genome-wide analysis of epistasis study, this first stage

20   involving the difference of correlation coefficient evaluations can be completed within

21   roughly 24 hours on commercially available GPU set-up compared to roughly a year on a

22   single-core CPU. From the subset of interactions deemed significant in the rapid filtering

1   stage, a second stage analysis is performed using the likelihood ratio statistical test on the

2   logistic regression to obtain the p-value on the estimated coefficients corresponding to the

3   intercept, individual effects of the single loci and the interaction terms. As this necessitates

4   only a minor amount of computations of logistic regressions in R using 'anova' test on the

5   'glm' fit with the 'binary' family option, for a retention rate of 6.8×10^{-6} , an expected

6   8.5×10^6 pairs, this requires some 2.5 days on a single core system of the hardware

7   specifications listed in the methods section in R. This is impractical, however, if we are to

8   limit ourselves to a range of top significant pairs which can be below a more stringent

9   threshold, e.g. 1.0×10^{-8}, it drops down to an expected number of some 600-700 pairs,

10   which requires around 150 seconds (4 computations per second) to validate. It should be

11   noted that dedicated software, such as INTERSNP [7] is considerably faster for this second

12   pass than pure R, the quoted figure of 8.5×10^6 interaction pairs should be done between

13   one or two hours using INTERSNP. A complete genome-wide association analysis with

14   INTERSNP on a single core would be in the order of a year. FASTEPISTASI.S would

15   have been some 70 days on a single core. Note that INTERSNP is quoted here for a full

16   logistic regression, whilst FASTEPISTASIS does a linear regression. Of course the

17   performance of both INTERSNP (which again is about two orders of magnitude faster than

18   plain R [using the glm() function]) and FASTEPISTASIS can be easily improved using

19   multi-core systems and clusters.

20   Including more SNPs into the second stage is feasible, of course. We have found a

21   threshold of 6.8 x 10-6 practical, lowering this by e.g. one order of magnitude will incur

22   only a slight increase in run-time for stage 1 and a linear increase for stage 2. Of course, if

1   the threshold for entry into stage 2 is lowered too much, hardware specifics such as disk

2   speed is becoming an issue in the performance of the program.

3   The reasoning behind the two-stage approach is threefold. First, the computations involved

4   in the first stage are much less extensive as compared to estimating for significance in

5   logistic regression. Second, a readily available R package, 'gputools', allows the estimation

6   of correlation coefficients to be performed on the graphic card which greatly reduces the

7   time and cost. Third, contrary to common multistage practice where the single locus test is

8   performed initially followed by higher order testing on loci which showed single locus

9   significance, the necessity of interaction loci to first show significant marginal effects is not

10 imposed, thus rendering this method a truly exhaustive search across all two-way

11 interactions. The results from the MS and Panic disorder analyses are used as preliminary

12 basis where this statement can be founded. A Plink method to test for univariate SNP

13 significance is used to provide an indication of the SNPs that would be kept using the more

14 traditional mandatory main effect significance. Firstly, referring to Tables S.2 and S.3 in

15 the supplemental section, it is shown that a vast majority of significant interaction pairs

16 would not have been captured if one is to pre-filter based on univariate significance.

17 Furthermore, referring to Figures S.2 to S.5 in the supplemental section, univariate p-values

18 are plotted against the interaction pairs captured by EPIBLASTER, the lack of trends helps

19 support that the method is indeed conducting the search unbiased to the marginal effects at

20 the two loci. High overestimation of the significance of the pair in the premilinary step

21 one filtering stage can occur when the SNPs are very rare. Severe underestimating of the

22 p-values using this approximation (false negatives) has also rarely been noticed but was

1   traced to a small subset of those SNP pairs that are in high linkage disequilibrium, which

2   are not the main focus of this method. For computational ease, no lower bound on physical

3   distance between SNPs or on LD between SNPs is imposed.

4   We also noted no inflation of the test statistic in our datasets, however, in cases it might be

5   advisable to include MDS por PCA components in the analysis,   e.g. by working on

6   residuals of the SNP genotypes on these components.

7

8   Overall, comparing the p-values obtained from FastEpistasis to the approximated p-

9   values tabulated from EPIBLASTER stage 1 show that although discrepancy in the p-

10   values do exist, the adopted method does manage to capture all of the significant pairs and

11   the occurrence of significant pairs being omitted is practically nil when the threshold p-

12   values is chosen to be far enough from the Bonferroni corrected global significance. Still

13   the computational load for the second stage analysis is negligible.

14   The concept of adopting the analysis of the difference of case-only and control-only

15   studies into an unified test has been suggested in prior studies analyze pairwise SNPs. Hoh

16   and Ott [8] initially proposed taking the ratios of the Chi-squares of the 3x3 contingency

17   tables between cases and controls as a measure of significance. Zhao et al.[9] and Zaykin

18   et al[10] have also proposed examining the gene interactions with a defined Linkage

19   Disequilibrium created by the interaction between two unliked loci. Significance is

20   evaluated with the analysis of the difference of the LD values between case-only and

21   control-only populations. Hardy-Weinberg Equilibrium must hold for this measure of

22   interaction and test statistics to be valid. Zhao et al. has further suggested the method

1   exhibits greater power than conventional linear regression, as it does not treat the

2   interaction as a residual term and allows for implicit nonlinear interaction, and faster

3   computational time than the traditional four degrees of freedom logistic regression model

4   rendering it more suitable for GWAS. The proposed method in this paper performs the

5   search in the first stage for only the effects of the interaction term by analyzing the

6   difference of the correlation coefficients as an indication for significance and then adopts

7   the more conventional logistic regression method to substantiate the findings on a subset of

8   pairs. As the difference is based on two separate groups, population stratification can have

9   an effect on the power of the method. However, looking at the number of pairs retained

10 from our examples, the actual inflation is very low. In the multiple sclerosis analysis,

11 423680 pairs is expected be below the 1×10^{-5} threshold, an observed number of pairs

12 captured is noted as 407660. The method can indeed be simplified to a case-only study, by

13 making the assumption that the correlation coefficient of the controls be null for all pairs,

14 this approach would further speed up the computational time by a factor of 2 at the expense

15 of potentially losing both potentially power and clearly precision. Moreover, the

16 approximation approach does not only apply to the dosage coding (0, 1, 2), but also to

17 other coding such as dominance, recessivity and heterozygosity. In general a p-value

18 cutoff of less than 1e-5 should indeed be sufficient to capture all the results with a p <1e-08

19 in the logistic regression and is, with all caution, suggested as a cut-off to be used in a first

20 analysis, truly making EPIBLASTER exhaustive within this setting.

21   With respect to the results from MS and panic disorder presented we note that, although,

22 there is no pair beyond a Bonferroni corrected threshold for significance at a corrected p-

value of 0.05, the marginal effects in the top 10 pairs do not at all show a tendency to deviate from a uniform distribution. This means that prefiltering pairs of SNPs on marginal p-values for subsequent epistasis analysis may be a less promising strategy than sometimes considered, although more analyses and larger sample sizes will be needed for a better founded statement on this issue.

In the editing phase of this article, it has come to our attention that Hu et al.[11] has also developed a strategy involving GPU to enhance genome-wide significant SNP pair interaction search, quoting a total runtime of 27 hours to scan through the Wellcome Trust Case Control Consortium's bipolar disorder data consisting of 500k SNPs. The proposed algorithm by Hu et al. helps consolidate the improved time performance using the inherent parallel nature of GPU to search for significance in all possible SNP pairs. The method is distinct from ours as it uses the a difference of odds ratios measure between cases and controls to pick significant SNP pair candidates.

We would like to point out that with EPIBLASTER it is possible to perform genome-wide analysis of epistasis on very small-scale and inexpensive hardware, reducing the need for large clusters for this kind of application.

Future work is planned to incorporate the logistic regression and other more novel definitions of gene-gene interactions onto the graphical processing units. EPIBLASTER is available at http:/www.mpipsykl.mpg.de/epiblaster.

Acknowledgements

the NGFN (Moods - 01GS08145 to BMM) and the project Control-MS within the "

Krankheitsbezogenes Kompetenznetz Multiple Sklerose is gratefully acknowledged

Conflict of Interest Statement

The authors declare no conflict of interest.

Supplementary information is available at European Journal of Human Genetics'

website

References

1 Marchini J, Donelly P, Cardon L.R.: Genome-wide strategies for detecting multiple loci

that influence complex diseases. Nature Genetics 2005; **37:** 413-417.

2 Gretton A, Borgwardt K, Rasch B, Schölkopf B, Smola A. A Kernel Method for the Two-

Sample-Problem. NIPS 2006: 513-520.

3 Wellek S, Ziegler A, : A Genotype-based approach to assessing the association between

single nucleotide polymorphisms. Human Heredity 2009; **67:** 128-139.

4 Erhardt A et al.: TMEM132D, a new candidate for anxiety phenotypes: evidence from

human and mouse studies. Molecular Psychiatry 2010 April [Epub ahead of print].

5 Nischwitz S et al.:Evidence for VAV2 and ZNF433 as susceptibility genes for multiple

sclerosis. Journal of Neuroimmunology 2010 June [Epub ahead of print].

6 Schüpbach T, Xenarios I, Bergmann S and Kapur K: FastEpistasis: a high performance

computing solution for quantitative trait epistasis. Bioinformatics 2010 **26**(11):1468-1469

7 Herold C, Steffens M, Brockschmidt F, Baur MP and Becker T: INTERSNP: genome-

wide interaction analysis guided by a priori information. Bioinformatics 2009;

25(24):3275-3281

8 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait

genes. Nature Reviews Genetics 2003; **4:** 701-709.

9 Zhao J, Xiong M: Test for interaction between two unlinked loci. The American Journal

of Human Genetics 2006; **79:** 831-845.

10 Zaykin DV, Meng Z, Ehm MG: Contrasting linkage-disequilibrium patterns between

cases and controls as a novel association-mapping method. The American Journal of

Human Genetics 2006; **78(5):** 737-46.

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 Research 2010 Jul; **20**(7):854-7.

1

2

3

4

5

6

7

8 **Figures, Equations and Tables.**

9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

1 <u>Equations</u>
2
3 Equation 1. Correlation coefficients (Pearson) difference between case-only and control-
4 only for each SNP-SNP pair. Note that no assumptions, such as HWE to hold, are needed
5 here.
6
7 Difference of Correlation Coefficients $= \Delta =$

8

$$\sum_{i=cases-only} \left\{ \frac{\left(SNP1_i - \overline{SNP1_i}\right)\left(SNP2_i - \overline{SNP2_i}\right)}{\left(n_i-1\right)\sigma_{SNP1_i}\sigma_{SNP2_i}} \right\} - \sum_{j=controls-only} \left\{ \frac{\left(SNP1_j - \overline{SNP1_j}\right)\left(SNP2_j - \overline{SNP2_j}\right)}{\left(n_i-1\right)\sigma_{SNP1_j}\sigma_{SNP2_j}} \right\}$$

10 The variance of each one of these correlation coefficients is, as shown by Wellek and
11 Ziegler [3], equal to $1/n-1$ where n is the respective number of cases and controls. As
12 the cases and controls constitute, obviously, independent samples, the total variance V_{tot} is
13 then the sum of the two single variances. As a consequence and from both Gretton et al[2]
14 and Wellek and Ziegler[3] thus $T = \Delta / V_{tot} \sim N(0,1)$.
15
16
17
18
19 Equation 2. Full rank logistic regression model.
20
21 Phenotype $=$ Intercept $+ \alpha SNP1 + \beta SNP2 + \gamma SNP1*SNP2$
22

23

<u>Titles and legends to figures.</u>

Figure 1. Histogram of differences of correlation coefficients of all 2-way interactions of 2000 SNPs exhibiting the expected Gaussian distribution shape.

Figure 2. Logarithmic p-values from interaction term of logistic regression versus correlation coefficient differences of all 2-way interactions from 2000 SNPs.

Figure 3. Logarithmic p-values from interaction term of the logistic regression model versus correlation coefficient differences p-values from 2000 SNPs (2000C2 = 1999000 SNP-SNP pairs). Quality of Fit (R2) between the p-values is 99.9%.

Figure 4. Panic disorder logarithmic p-values density plot: Top 10 SNP pairs (points marked in black) and threshold correlation coefficient difference pvalue. FastEpistasis p-values are on the y-axis, p-values from EPIBLASTER on the x-axis.

Figure 5. Panic disorder logarithmic p-values density plot: Top 100 SNP pairs (points marked in black) and threshold correlation coefficient difference p-value. FastEpistasis p-values are on the y-axis, p-values from EPIBLASTER on the x-axis.

Figure 6. Multiple sclerosis logarithmic p-values density plot: Top 10 SNP pairs (points marked in black) and threshold correlation coefficient difference p-value

Figure 7. Multiple sclerosis logarithmic p-values density plot: Top 100 SNP pairs (points marked in black) and threshold correlation coefficient difference p-value.

Table 1. Top 10 panic disorder SNP pairs difference of correlation coefficient, FastEpistasis and logistic regression p-values.

Ranking	SNP1 Name	SNP2 Name	Diff. of R	Diff. Of R Pvalue	FastEpistatic Pvalue	Lreg-SNP1 Pvalue	Lreg-SNP2 Pvalue	Lreg-Interaction Pvalue
Top 10 Panic Disorder SNP pairs ranked by FastEpistasis								
1	rs4653309	rs17338700	0.61	2.26E-10	6.97E-12	1.94E-01	6.16E-01	8.98E-12
2	rs4984422	rs1967113	-0.66	8.86E-12	1.06E-11	2.02E-01	8.66E-01	4.72E-13
3	rs1156847	rs7246846	0.62	1.21E-10	6.57E-11	7.45E-01	2.01E-01	4.02E-11
4	rs6455842	rs265548	0.64	2.81E-11	7.25E-11	3.67E-01	3.55E-01	6.17E-12
5	rs12188192	rs1317584	0.58	1.98E-09	8.08E-11	2.51E-01	4.90E-01	1.87E-10
6	rs2100807	rs4875302	0.61	1.95E-10	1.88E-10	7.15E-01	3.95E-01	1.45E-10
7	rs11900448	rs11939830	-0.6	5.07E-10	2.17E-10	2.27E-01	7.97E-01	2.64E-10
8	rs6762261	rs4745430	0.61	3.24E-10	2.36E-10	4.81E-01	1.87E-01	1.65E-10
9	rs2374344	rs1011308	0.55	1.26E-08	2.53E-10	3.97E-01	4.00E-01	4.41E-10
10	rs11925795	rs4731772	0.62	1.60E-10	3.11E-10	3.49E-03	4.37E-01	4.71E-10

1
2

4
5
6
7
8
9
10
11
12
13

Table 2. Top panic disorder SNP pairs annotations.

Top10 Panic Disorder SNP pairs ranked by FastEpistasis Annotations							
SNP1 Name	Chromosome	Basepair	Gene	SNP 2 Name	Chromosome	Basepair	Gene
rs4653309	chr1	37876927		rs17338700	chr2	33841677	
rs4984422	chr15	94456392		rs1967113	chr18	26830011	DSC3
rs1156847	chr9	2586783		rs7246846	chr19	56705171	
rs6455842	chr6	162962566	PARK2	rs265548	chr19	17763334	
rs12188192	chr5	136380739	SPOCK1	rs1317584	chr6	12450775	
rs2100807	chr3	117506680	LSAMP	rs4875302	chr8	4028885	CSMD1
rs11900448	chr2	149650765	LOC130576	rs11939830	chr4	157150631	
rs6762261	chr3	136073828	EPHB1	rs4745430	chr9	77461845	
rs2374344	chr2	41994977		rs1011308	chr9	72478076	
rs11925795	chr3	178001610		rs4731772	chr7	130582931	

1

3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

1
2

Table 3. Top 10 multiple sclerosis SNP pairs difference of correlation coefficient, FastEpistasis and logistic regression p-values.

| Top 10 multiple sclerosis SNP pairs ranked by FastEpistasis							
Ranking	SNP1 Name	SNP2 Name	Diff. of R	Diff. Of R Pvalue	FastEpistatic Pvalue	Lreg-SNP1 Pvalue	Lreg-SNP2 Pvalue	Lreg-Interaction Pvalue
1	rs1392773	rs1384731	0.36	3.68E-11	3.78E-12	7.00E-01	9.90E-01	4.28E-12
2	rs1552621	rs6817936	0.34	2.00E-10	5.87E-11	7.85E-01	6.75E-01	6.71E-11
3	rs11710441	rs13226149	-0.33	5.92E-10	7.28E-11	4.79E-01	2.21E-01	8.66E-11
4	rs2218314	rs1384731	0.33	8.68E-10	8.19E-11	6.11E-01	9.99E-01	9.34E-11
5	rs6738313	rs3752735	-0.34	3.18E-10	1.03E-10	1.91E-01	7.18E-01	1.12E-10
6	rs7593466	rs11658318	0.34	3.14E-10	1.07E-10	7.66E-01	1.83E-01	1.01E-10
7	rs6758449	rs10055397	0.34	3.64E-10	1.09E-10	3.23E-01	6.90E-01	1.16E-10
8	rs17648731	rs7386137	0.35	1.22E-10	1.10E-10	8.15E-01	6.03E-01	1.05E-10
9	rs6550306	rs10503253	0.34	4.17E-10	1.52E-10	9.32E-01	3.58E-01	1.66E-10
10	rs2542509	rs2916433	0.34	5.04E-10	1.71E-10	6.94E-02	9.53E-01	1.81E-10

4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Table 4. Top 10 multiple sclerosis SNP pairs annotations.

| SNP1 Name	Chromosome	Basepair	Gene	SNP 2 Name	Chromosome	Basepair	Gene
rs1392773	chr4	143053312		rs1384731	chr5	10660797	
rs1552621	chr3	67460533		rs6817936	chr4	167934823	SPOCK3
rs11710441	chr3	145154009		rs13226149	chr7	94863536	PON3
rs2218314	chr4	143031581		rs1384731	chr5	10660797	
rs6738313	chr2	3382368	TTC15	rs3752735	chr18	49363018	
rs7593466	chr2	208807724	IDH1	rs11658318	chr17	27230172	UTP6
rs6758449	chr2	68290612	PPP3R1	rs10055397	chr5	120950796	
rs17648731	chr2	77575007		rs7386137	chr8	142596655	
rs6550306	chr3	34873129		rs10503253	chr8	4168252	CSMD1
rs2542509	chr2	71443251	ZNF638	rs2916433	chr4	4343724	LYAR/ZNF509

Top10 Multiple Sclerosis SNP pairs ranked by FastEpistasis Annotations

Simulation Data

Distribution of Correlation Coefficients Difference
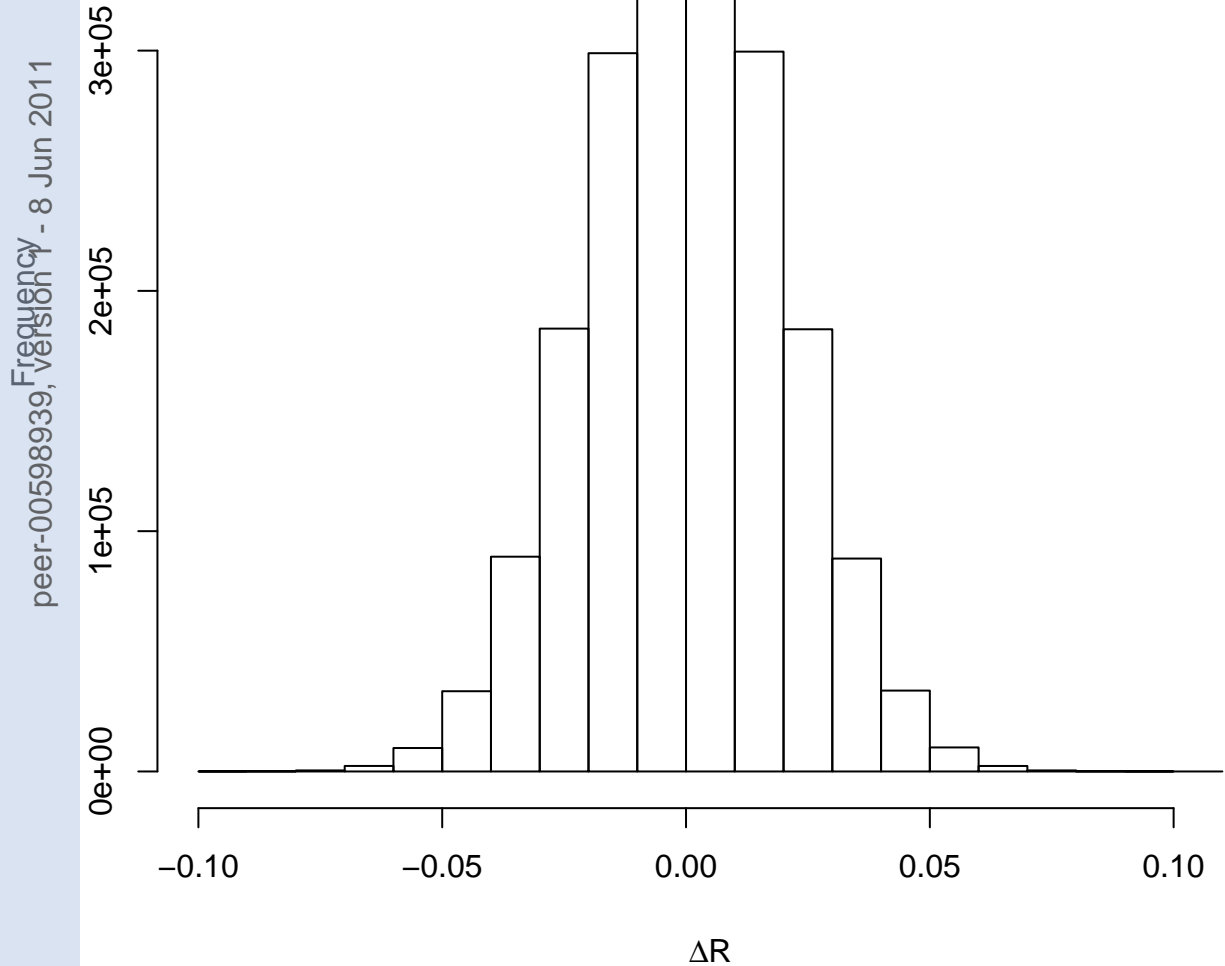
Simulation Data
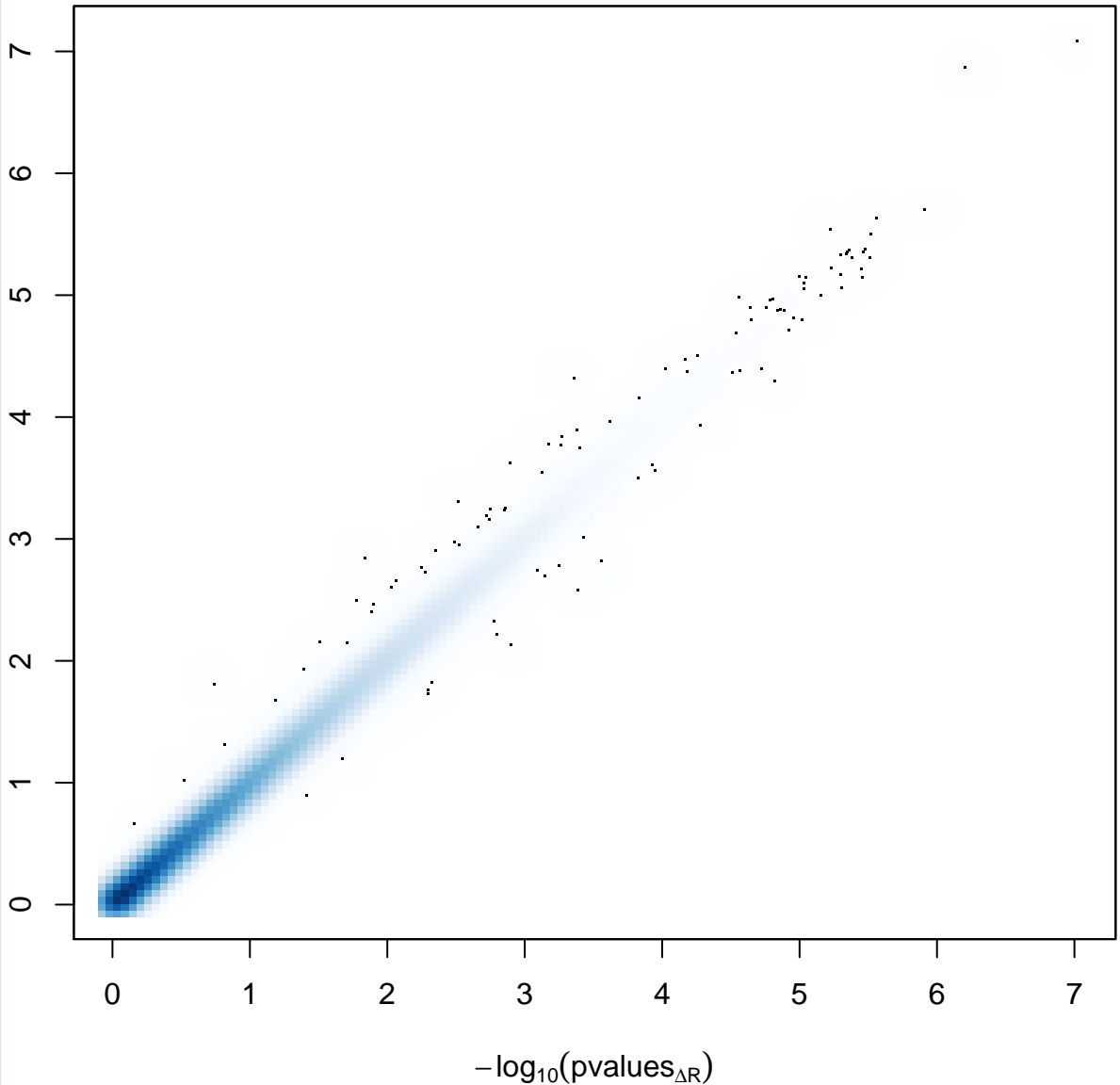### Logistic Regression p−values vs. Correlation Coefficients Difference

$-\log_{10}(\text{pvalues}_{\text{Logistic}})$

ΔR
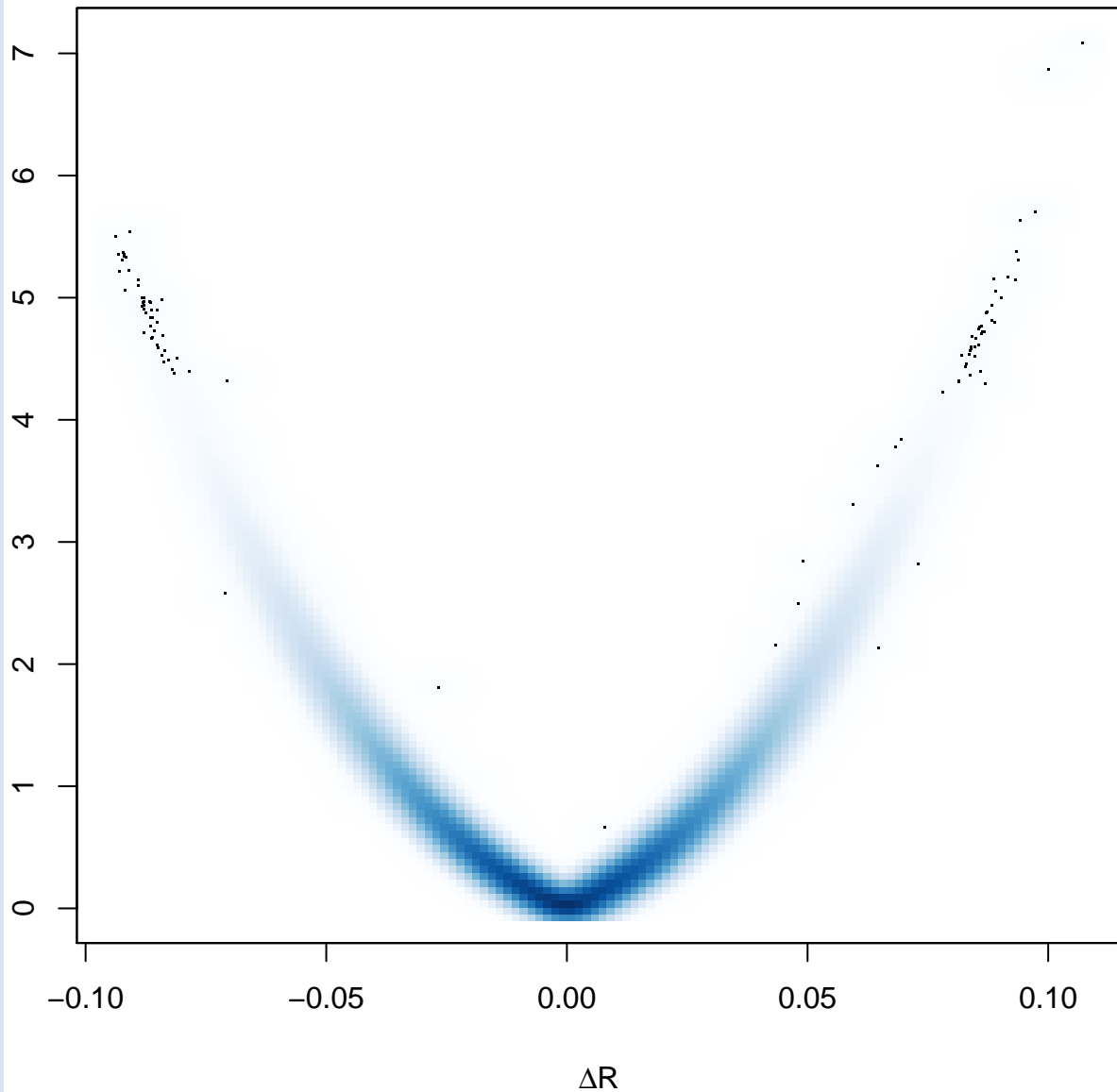
Simulation Data

Logistic Regression p−values vs. Correlation Coefficients Difference p−values

$-\log_{10}(pvalues_{\Delta R})$

Panic disorder

FastEpistasis versus correlation coefficient difference pvalues

Threshold 1.258e−08

$-\log_{10}(\text{pvalues}_{\text{FastEpistasis}})$

$-\log_{10}(\text{pvalues}_{\Delta R})$

Multiple Sclerosis

FastEpistasis versus correlation coefficient difference pvalues

Threshold 8.680e−10

$-\log_{10}\left(\text{pvalues}_{\text{FastEpistasis}}\right)$

$-\log_{10}\left(\text{pvalues}_{\Delta R}\right)$

Multiple Sclerosis
FastEpistasis versus correlation coefficient difference pvalues

Threshold 4.787e−07

$-\log_{10}(\text{pvalues}_{\Delta R})$