Identification of Genetic Interactions in Multi-Phenotype Studies

Max Planck Institute of Psychiatry
German Research Institute of Psychiatry

Beibei Jiang[1], Benno Pütz[1], Alessandro Gialluisi[2], Till Andlauer[1], NeuroDys Consortium, Nazanin Mirza-Schreiber[1], Bertram Müller-Myhsok[1].
[1]Max Planck Institute of Psychiatry, Munich, Germany
[2]Department of Epidemiology and Prevention, IRCCS Neuromed, Pozzilli, Italy

MAX-PLANCK-GESELLSCHAFT

1 Introduction

Genome-wide association studies (GWAS) are conducted to identify genetic variants associated with human complex traits. However, not all phenotypic variation can be explained by studying the effects of single genetic locus. Genetic interaction effects (epistasis) are one candidate mechanism for missing heritability. However, the search for significant epistasis (gene-gene interactions) poses as a computational challenge for modern day computing systems.

In the past five years, we have presented several efficient approaches to detect epistasis by exhaustive testing of all possible SNP pairs. The actual implementation of these approaches is done on the highly parallelized architecture available on Graphics Processing Units (GPU) rendering the completion of the full search feasible typically within one day.

Meanwhile, since several disease-associated polymorphisms have been identified by GWAS and an increasing number of phenotypes covering more information are available, we put our focus on studying epistasis in multi-trait studies and hope to link functional interactions between multiple traits, diseases and genetic factors (pairs of SNPs).

2 Motivation

In our previous epistasis study on the genome-wide sample of dyslexic children (age 11–19), three significant (Bonferroni corrected significant level: $p \leq 1.61{\times}10^{-12}$) interactions for the endophenotypes, word reading (WR), phoneme deletion (PD) and word spelling (WS) could be detected (Fig. 1).

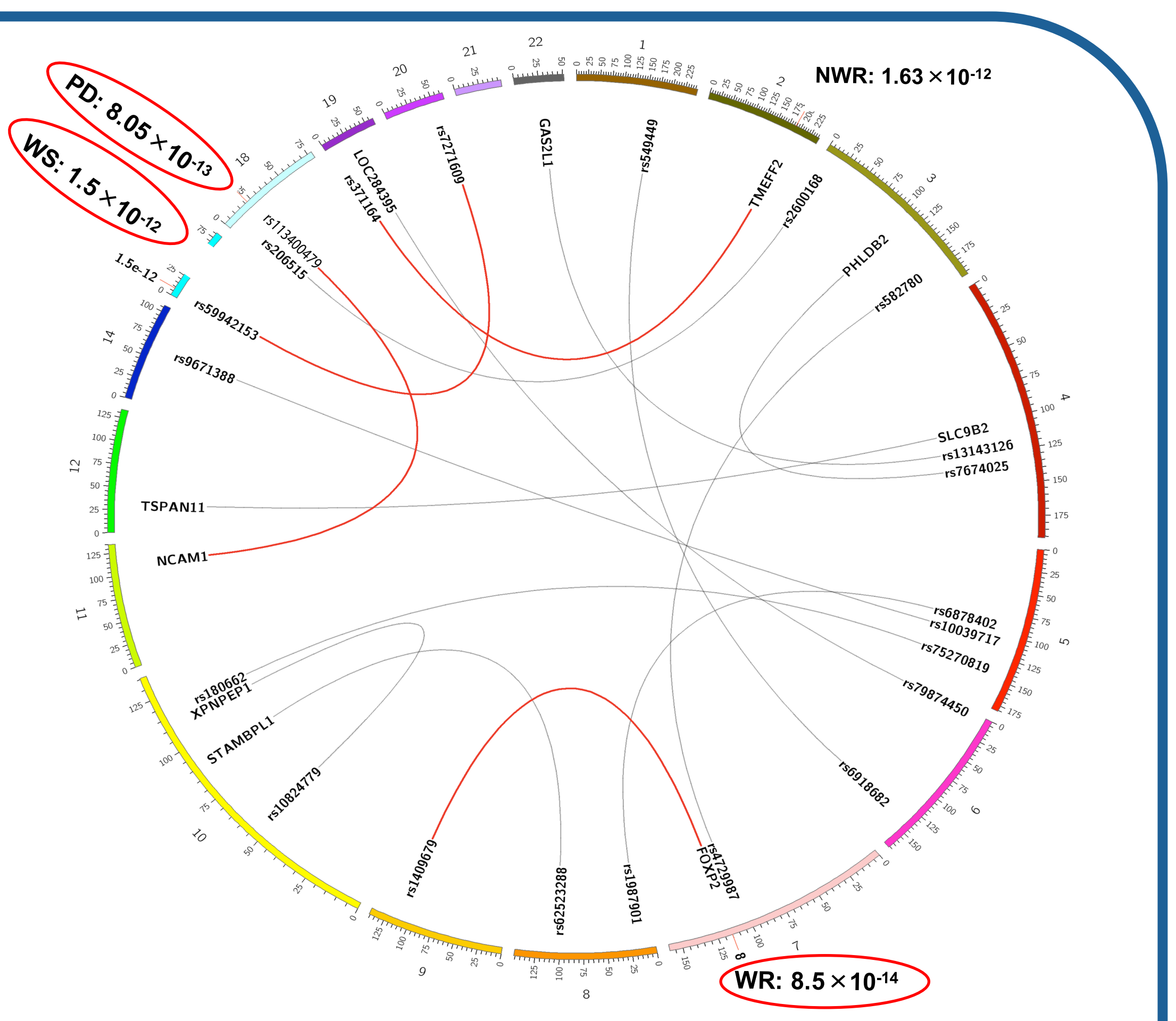


Fig.1: Inter-chromosomal SNP-SNP interactions detected for each endophenotype. Significant interactions are red-rimmed. Red and grey (*p ≤ 1×10*$^{-10}$) lines depict the chromosomal positions.

By conducting correlation tests on 10 phenotypes from the dyslexia study (~ 3500 individuals, ~ 6 million SNPs), strong relatedness was found between some phenotypes (Fig. 2).

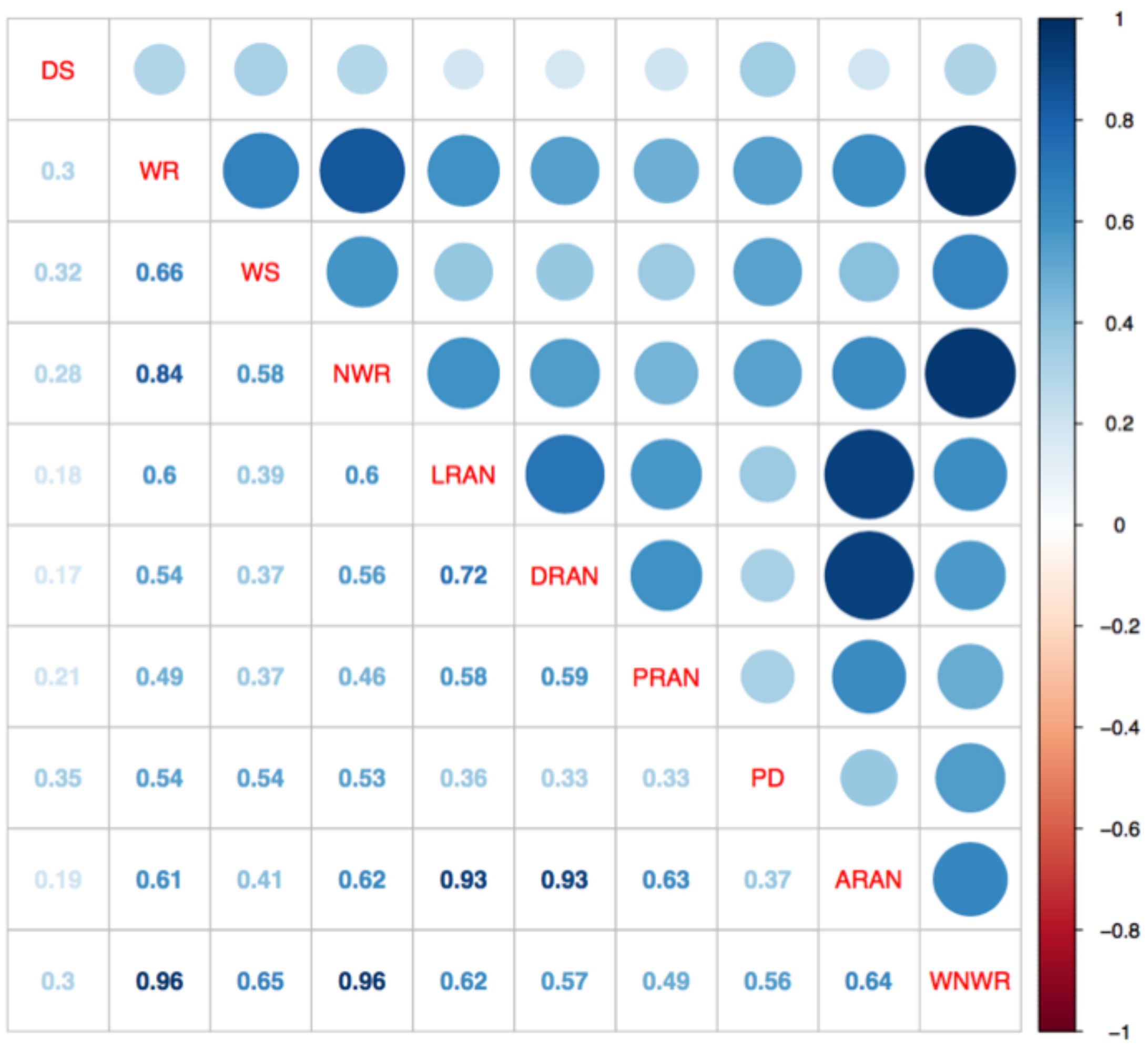


Fig.2: Correlation matrix of dyslexia phenotypes
DS: digit span; WR: word reading; WS: word spelling; NWR: non-word reading; LRAN: letters rapid automatic naming; DRAN: digits rapid automatic naming; PRAN: pictures rapid automatic naming; PD: phoneme deletion; ARAN: alphanumeric rapid automatized naming ; WNWR: word and non-word reading.

3 Methods

According to the definition of the Hilbert-Schmidt Independence Criterion (HSIC), $\text{Multi-Pheno-epiHSIC}_{\text{empirical}}((X,Y),\mathcal{F},\mathcal{G})$ can be computed in a runtime which is linear in n by rewriting it as

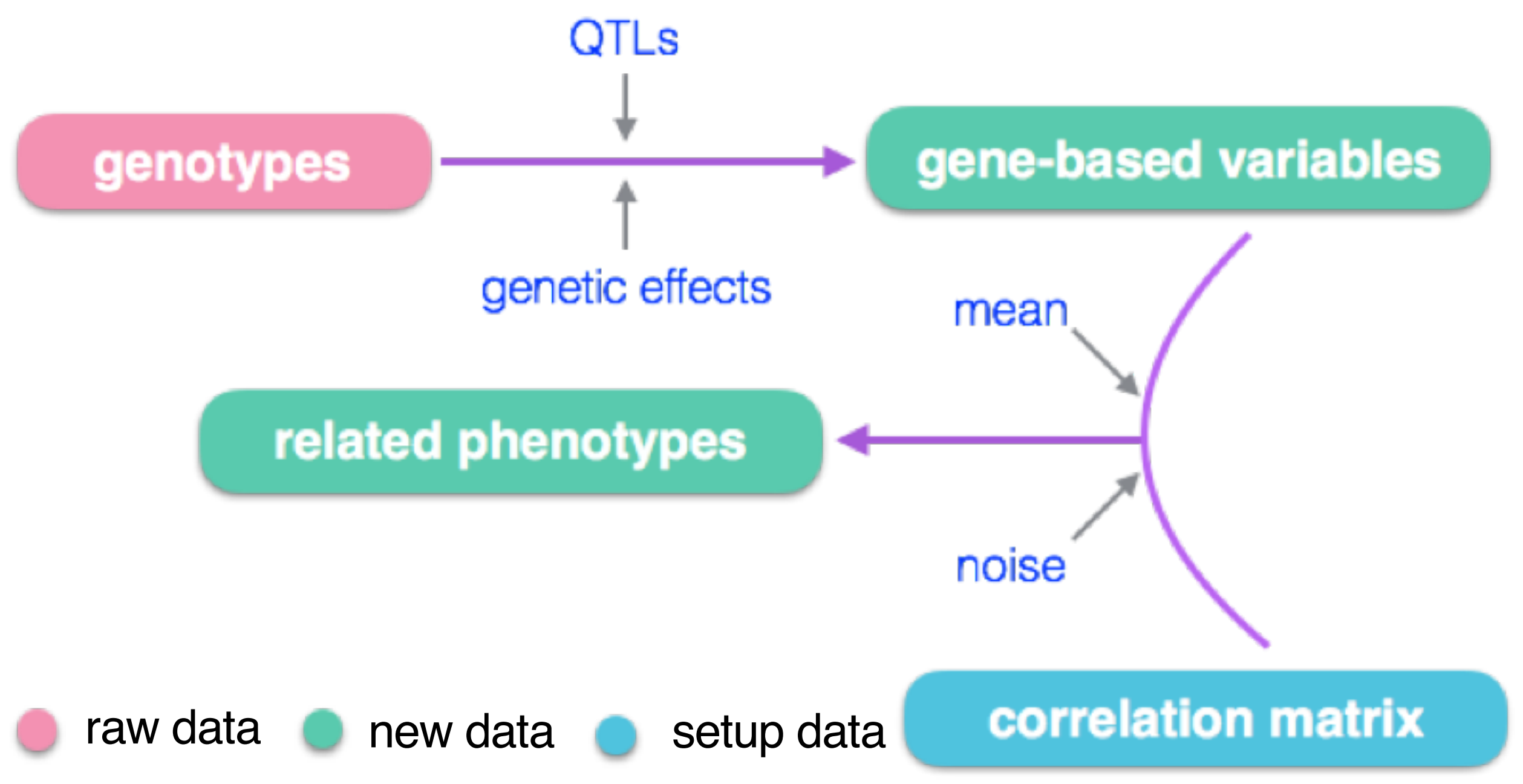
$$\text{Multi-Pheno-epiHSIC}_{\text{empirical}}((X,Y),\mathcal{F},\mathcal{G}) \propto \sum_{i=1}^{n} \tilde{x}_i^A \tilde{x}_i^B \prod_{j=1}^{m} \tilde{y}_i^j$$

X, *Y* are the random variables; $\mathcal{F},\mathcal{G}$ are the feature spaces on *X* and *Y*;
i and *j* are the indices of individuals and phenotypes (*i* = 1, 2, …, n; *j* = 1, 2, …, m), respectively;
$\tilde{x}_i^A$ and $\tilde{x}_i^B$ are SNPs which have been centered and divided by the standard deviation, respectively;
$\prod_{j=1}^{m} \tilde{y}_i^j$ is the product of multiple "phenotypes" which are the PCA components of the real phenotypes.

4 Results

Simulate multiple phenotypes with epistasis effects

R package on CRAN: *SimPhe*


QTLs
genotypes
gene-based variables
genetic effects
mean
related phenotypes
noise
correlation matrix
● raw data ● new data ● setup data

Permutation test

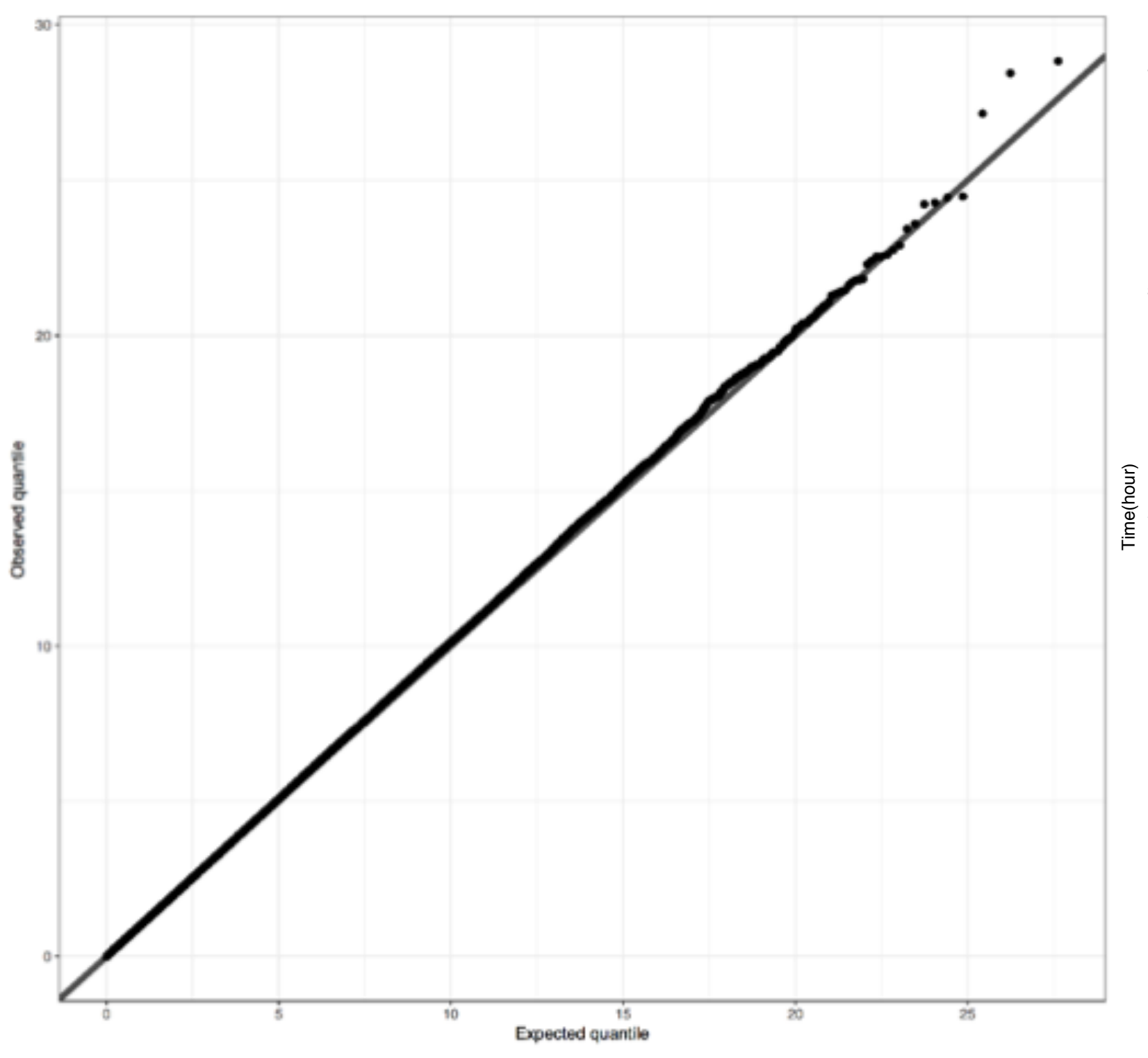


Fig.3: Q-Q plot of p-values from permutation test for analysis with simulated data. One million permution tests were conducted. X axis shows the theoretical values and Y axis shows the observed values. Both values were –log10 -transformed.

Fast GPGPU matrix manipulations

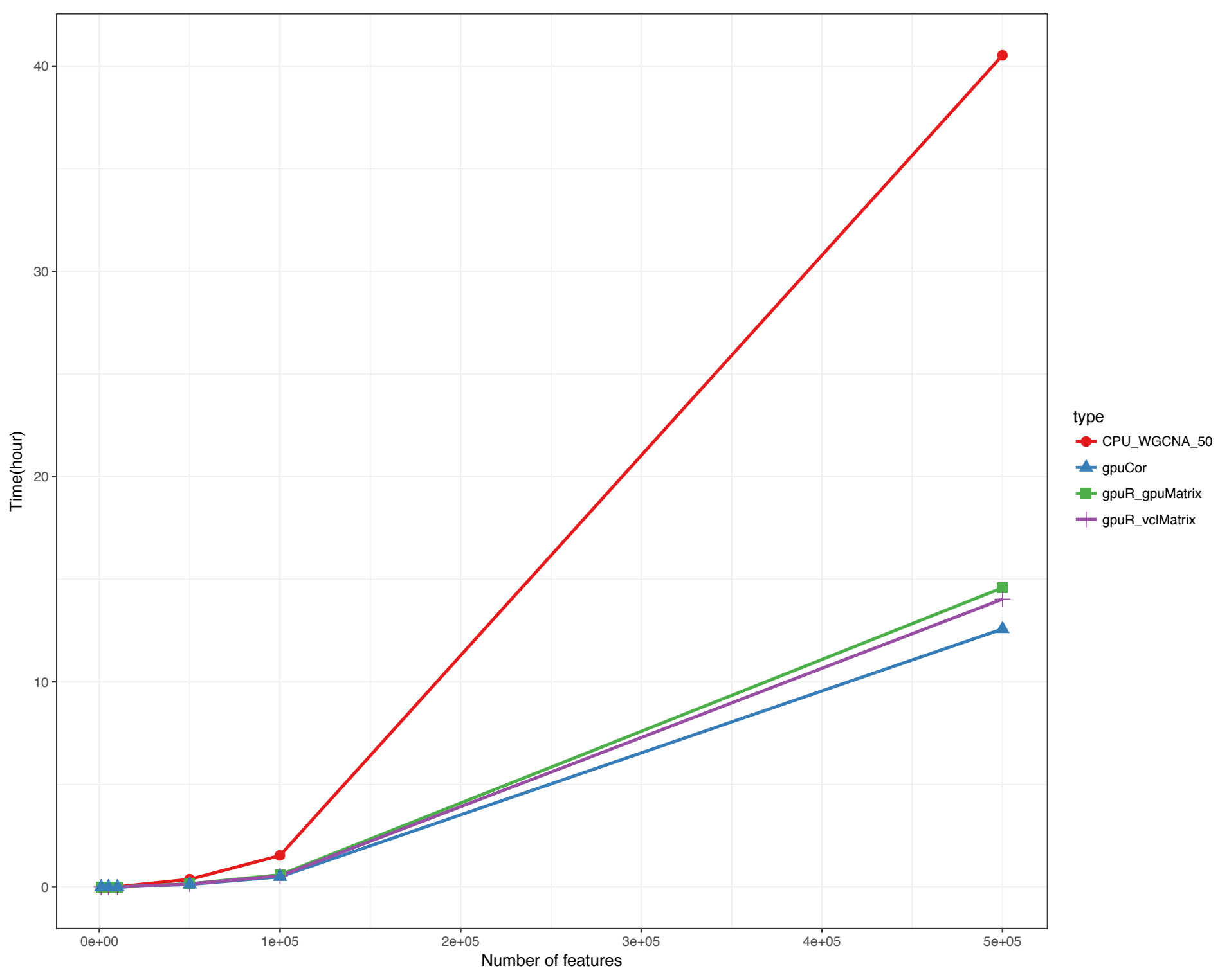


Fig.4: Computing time of core function on Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz and Tesla P100 (GPU) with 10000 individuals (~ 50 million SNP pairs) and different number of features (level: 1000, 5000, 10000, 50000, 100000, 500000).

5 Disscussion

Conclusions

- Simulation studies showed Multi-Pheno-epiHSIC method with dimentionality reduction of phenotypes can be applied to multiple-trait studies with precision and power.
- Fast GPGPU matrix manipulations ensable epistasis analysis in large datasets.

Outlook

- Apply Multi-Pheno-epiHSIC on dyslexia dataset to identify candidate variants which could not be explored by single-phenotype studies.
- Develop tools to deal with multi-omics datasets (more than two), hoping to understand more about the biological association / network.