Evaluation of epistasis detection methods for quantitative phenotypes

Authors: Stanislav Listopad[1], Gauri Renjith[2], and Qian Peng[1].

Author Affiliations:

[1]Department of Neuroscience, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA 92093, USA.

Corresponding Authors:

Stanislav Listopad: slistopad@scripps.edu

Qian Peng: qpeng@scripps.edu

Keywords: epistasis, genetic interaction, simulation, quantitative phenotype, software benchmark

Abstract

Background

Epistasis, or genetic interaction, has been increasingly recognized for its ubiquity and for its role in susceptibility to common human diseases, such as Alzheimer's. A wide variety of epistasis detection tools are currently available with several studies comparing the performance of methods suitable for case-control data. However, there is limited understanding of how well these tools perform with quantitative phenotypes.

Methods

We identified six epistasis detection methods suitable for quantitative phenotype data: EpiSNP, Matrix Epistasis, MIDESP, PLINK Epistasis, QMDR, and REMMA. To evaluate these tools, we generated simulated datasets using EpiGEN. The datasets modeled various pairwise interactions between disease-associated SNPs, including dominant, multiplicative, recessive, and XOR interactions. Additionally, we assessed the BOOST and MDR algorithms on discretized (case-control) version of the datasets. These tools were then tested on the Adolescent Brain Cognitive Development (ABCD) dataset for the externalizing behavior phenotype.

Results

Each tool exhibited strong performance for certain interaction types, but weaker performance for others. MDR achieved the highest overall detection rate of 60%, while EpiSNP had the lowest overall detection rate of 7%. MDR and MIDESP performed best at detecting multiplicative interactions with detection rates of 54% and 41% respectively. Both MDR and MIDESP were also effective at detecting XOR interactions with detection rates of 84% and 50% respectively. PLINK Epistasis, Matrix Epistasis, and REMMA excelled at detecting dominant interactions, all achieving a 100% detection rate. On the other hand, EpiSNP was particularly

effective at detecting recessive interactions with a detection rate of 66%. When analyzing the ABCD dataset, Plink Epistasis and Plink BOOST identified SNPs within the *DRD2* and *DRD4* genes, which have been previously linked to externalizing behavior.

Conclusion

Since no single method consistently outperforms others across all types of epistasis, and given that the specific types of epistasis present in a dataset are often unknown, it may be more effective to use multiple epistasis detection algorithms in combination to obtain comprehensive results.

Introduction

Over a decade of genome-wide association studies (GWAS) have generated robust results for a broad range of complex traits and diseases with increasingly large sample sizes [1], [2]. While GWAS is a powerful tool for genetic mapping, most variants have shown small effects and captured a fraction of the disease heritability in general. The gap between the phenotypic variations that can be explained by genetic variants identified by GWAS and that were estimated by twin and family studies is referred to as “missing heritability” [3][4], [5]. A number of factors have been hypothesized to account for missing heritability. Genetic interaction is one such potentially critical culprit [6]. The present association studies model genetic effects of multiple variants in a linear combination (assuming additivity), while combinatorial effects beyond the linear effects are not considered. When phenotypic effects from two genetic variants deviate from the expected additive value of the individual mutations, the two variants are said to exhibit an epistatic interaction [7]. Once treated as an exception, epistasis has been increasingly recognized as ubiquitous [8], [9]. It has been shown that disease genes have a high propensity to

interact with each other [10]. Additionally, the majority of genetic variants associated with complex diseases are located in non-coding regions, suggesting that they may influence traits through regulatory interactions [11].

Although epistasis has not been systematically considered in most genetic studies, evidence suggests that it contributes to susceptibility to common human diseases, such as Alzheimer's [12], [13][14], [15]. The term epistasis may refer to biological or statistical epistasis [16], [17]. Biological epistasis involves physical interaction between two or more biological components. For instance, hair color in humans is influenced by such epistasis: the interaction between *MC1R* and *ASIP* genes is partially responsible for red hair color [18]. Meanwhile, statistical epistasis refers to the departure from additive effects of genetic variants at different loci with regard to their global contribution to the phenotype. Biological epistasis can be modeled by statistical epistasis, although statistical interaction does not always imply biological interaction. Elucidating biological interaction from statistical one is a challenging task. Improving the quality of statistical epistasis detection is nonetheless a crucial step to enable the discovery of more biological interactions. These biological interactions, may then in turn, improve our understanding of heritable traits. Given the availability of large GWAS datasets and a number of epistasis detection methods that have been developed over the years, statistical epistasis analyses may be potentially carried out at large scales. Several publications have compared the performance of epistasis detection methods primarily suitable for case-control data [19], [20], [21], [22], [23]. However, less is known about the performance of epistasis detection tools suitable for quantitative phenotypes. To this end, we conducted a survey of the commonly used epistasis detection tools, evaluating their performance on a set of simulated data, followed by testing on a real-world dataset.

For this study, we used EpiGEN to generate simulated epistasis datasets with quantitative phenotypes [24]. Given the combinatorial nature of genetic interactions, many types of epistasis theoretically exist. We modeled four major types of interactions that are most plausible: dominant, multiplicative, recessive, and exclusive-or (XOR). In the dominant model, as defined by EpiGEN, an interaction occurs if both single nucleotide polymorphisms (SNPs) each have at least one minor allele. In the multiplicative model, an interaction occurs if either SNP carries at least one minor allele, with the strength of the interaction increasing proportionally to the number of minor alleles present in the interacting SNPs. In the recessive model, an interaction is observed only if both SNPs have two minor alleles. In the XOR model, an interaction occurs only if one of the SNPs, has at least one minor allele, but not both. Note that other definitions of XOR epistasis exist, such as the one defined by Sha [25]. While the dominant, recessive, and XOR interactions are considered biologically plausible [25], [26], [27]. There is no clear evidence establishing the existence of multiplicative interactions, to the best of our knowledge. However, the multiplicative interaction is often used as an approximation for various types of interactions, especially when the exact nature of the interactions is uncertain. Throughout this publication, we use the EpiGEN-specific definitions of dominant, multiplicative, recessive, and XOR epistasis, as these terms may have alternative meanings in broader genetic contexts.

After generating the simulated data, we identified several publicly available epistasis detection tools suitable for analyzing quantitative phenotype data, including EpiSNP, Matrix Epistasis, MIDESP, PLINK Epistasis, QMDR, and REMMA [28], [29], [30], [31], [32], [33] . Our primary goal was to examine tools that could directly analyze quantitative phenotype data, which, when available, usually offers increased statistical power for detection. However, we also acknowledge that in some cases the phenotype could be discretized into a binary (case-control)

value, which would allow us to utilize a wider range of analysis methods. Therefore, we also evaluated the Boolean operation-based screening and testing (BOOST) algorithm as implemented in PLINK and the multifactor dimensionality reduction (MDR) algorithm as implemented in QMDR with discretized version of our simulated phenotype [34], [35]. A complete list of tools we attempted to evaluate is provided in Table 1. While our coverage of tools is not exhaustive, we aimed to include tools that employ some of the most commonly used epistasis detection models, such as linear regression and linear mixed models, among others. We then assessed the performance of each tool in detecting dominant, recessive, multiplicative, and XOR interactions.

Epistasis can involve interactions between two or more SNPs, but this review focuses exclusively on pairwise interactions. Interaction between two SNPs is referred to as second order epistasis, while interaction between more than two SNPs is referred to as higher order epistasis. There are two categories of epistasis detection methods: exhaustive and non-exhaustive. Exhaustive algorithms test every possible combination of SNPs, which can quickly become unsustainable for higher order epistasis as the number of comparisons grows exponentially with the order of epistasis. The huge number of comparisons then results in lengthy runtime and low statistical power. Non-exhaustive epistasis detection methods mitigate these challenges by evaluating only a subset of all possible comparisons. However, there is evidence suggesting that they may perform worse than exhaustive methods [21]. For these reasons, this study focuses on 2nd order epistasis detection methods.

In addition to using simulated data, we evaluated the tools in this study using the Adolescent Brain Cognitive Development (ABCD) dataset. Unlike simulated data, ABCD dataset includes population structure, individual relatedness, and multiple covariates, which add

complexity to the analysis. It also contains a larger number of samples and SNPs compared to our simulated datasets. Therefore, the ABCD dataset provided a valuable opportunity to compare the performance of epistasis detection tools in a real-world context. For this study, we focused on the externalizing behavior phenotype, which is both heritable and common [36].

Name	Underlying Model	Simulated Data	ABCD Data
EpiSNP [29]	General Linear Model	✓	Errors during execution.
FastANOVA [37]	ANOVA	Failed installation.	Not Applicable
FRGEpistasis [38]	Functional Regression Model	Can only perform gene x gene interaction testing, not SNP x SNP.	Not Applicable
Matrix Epistasis [39]	Linear Regression	✓	✓
MIDESP [31]	Mutual Information	✓	✓
PLINK (BOOST and Epistasis) [32], [34]	BOOST, Linear Regression	✓	✓
QMDR [33], [35]	MDR	✓	✓
REMMA [28]	Linear Mixed Model	✓	✓
SNPassoc [40]	Log Likelihood Ratio Test	Successfully used with <200 SNPs, but could not use with 1000 SNPs.	Not Applicable
WISH-R [41]	Linear Regression	Failed Installation.	Not Applicable

Table 1: List of tools evaluated, their underlying epistasis models, and their evaluation statuses. *Underlined tools were further evaluated in the study. Checkmarks (✓) indicate successful installation and usage.*

Methods

Data Generation

Quantitative Phenotype Datasets

A total of 40 datasets were generated using EpiGEN [24], each containing a single type of epistatic interaction: dominant, recessive, multiplicative, or XOR. Within each dataset, all interactions had the same interaction alpha [24]. Interaction alpha is a positive number that quantifies interaction strength. Definitions of dominant, recessive, and multiplicative interactions in terms of interaction alpha can be found in the EpiGEN publication. The XOR was defined as follows: if one SNP genotype is in {1, 2} (at least one minor allele) and the other SNP genotype is in {0} (no minor alleles) then there is an interaction. There are 20 multiplicative, 8 recessive, 8 dominant, and 4 XOR datasets. All interactions occurred between pairs of disease SNPs. Each dataset contained 1000 SNPs and 1000 individuals, with the data generated using chromosome 1 of the CEU HapMap3 cohort [42].

Out of the 20 multiplicative datasets, 16 contained 5 disease SNPs, and 4 contained 20 disease SNPs. The datasets with 5 disease SNPs were created by permuting the following variables: interaction alpha {1.25, 1.5, 2, 3}, number of interacting SNP pairs {1, 2}, and purity status {pure, impure}, resulting in 16 datasets. Lower interaction alphas such as 1.25 and 1.5 produce phenotype values for individuals with interacting disease SNPs that are just a few times larger than those without interacting disease SNPs, making detection challenging. Larger interaction alphas such as 2 and 3 lead to phenotype values for individuals with interacting disease SNPs to be tens of times larger than phenotype values of individuals without interacting disease SNPs, making these interactions easier to detect. Pure datasets contained no individual main effects, while impure datasets included some individual main effects [19]. The 4 datasets with 20 disease SNPs were created by permuting interaction alphas {2, 3} and purity status

159 {pure, impure}, with each dataset containing 8 interacting disease SNP pairs. In total, the 20
160 multiplicative datasets contained 56 interacting SNP pairs (8 * 1 + 8 * 2 + 4 * 8 = 56).

161 The 8 dominant and 8 recessive datasets each contained 5 disease SNPs. These datasets
162 were created by permuting interaction alpha values {8, 16}, the number of interacting SNP pairs
163 {1, 2}, and purity status {pure, impure}. The 8 dominant and 8 recessive datasets contained a
164 total of 24 interacting SNP pairs (4 * 1 + 4 * 2 + 4 * 1 + 4 * 2 = 24). The four XOR datasets each
165 contained 20 disease SNPs with 8 interacting SNP pairs, resulting in a total of 32 interacting SNP
166 pairs across the XOR datasets. These datasets were created by permuting interaction alphas {8,
167 16} and purity status {pure, impure}. Interaction alphas of 8 and 16 were chosen to ensure that
168 the phenotypes of individuals with interacting disease SNPs were distinct from those without
169 interacting disease SNPs. Full in-depth specifications used to generate each dataset can be found
170 in supplemental file S1.

171 **Case-Control Datasets**

172 All of the above 40 datasets were also discretized to create an additional set of 40 case-
173 control datasets. The discretization thresholds were selected individually for each dataset to
174 ensure a reasonable balance of cases and controls, typically resulting in at least 1% of samples
175 being cases. While discretizing the phenotype results in a loss of information, it enables the use
176 of a much larger number of tools that are suitable for case-control epistasis analysis, compared to
177 tools designed for quantitative epistasis analysis. Further details on the discretization process can
178 be found in supplemental file S1.

179 **Adolescent Brain Cognitive Development (ABCD) Dataset**

The ABCD study is an ongoing project aimed at studying brain development and child health [43]. One of the more highly heritable quantitative phenotypes in the study is externalizing behavior [44], which comprises of a wide range of antisocial behaviors such as aggression, bullying, conduct problems, rule breaking, and substance use [45], [46], [47]. Due to its common occurrence among children, externalizing behavior is an ideal phenotype for study. Additionally, some genetic interactions associated with externalizing behavior, such as DRD4-DRD2 and DRD4-SLC6A4 interactions, have already been identified [48], [49], [50]. DRD4 and DRD2 are dopamine receptor genes, while SLC6A4 is a serotonin transporter gene. The ABCD study contains various externalizing scores. For this analysis, we used the child behavioral check list (CBCL) externalizing t-score, with t-score > 64 indicating clinical symptoms [51], [52].

The cohort includes 11,666 children, of which 417 had externalizing t-score > 64, after excluding 202 individuals due to missing principal components of kinship matrix. The externalizing t-scores were normalized for age and sex, and residuals were used to further adjust for population stratification (using principal components of the kinship matrix) and batch site covariates (Figure 1) [53]. For REMMA, which already accounts for population stratification, only batch site covariate was used to obtain the residuals. While some of the tools in the study, such as Matrix Epistasis, can directly account for covariates, we opted to use residuals to ensure uniformity across all tools. The discretized phenotypes were obtained by converting the residuals into case-controls with a threshold of 20.4, chosen to match the ratio of cases to controls when using the original t-score threshold of 64. Genetic data was obtained using hg19 smokescreen array as documented in the ABCD study [54], with 515,279 SNPs in the dataset. Initially, we intended to perform epistasis analysis on the entire dataset, but only PLINK was able to complete the analysis within 24 hours. Therefore, we confined the analysis to SNPs on chromosome 11,

203 which includes 24,699 SNPs. Chromosome 11 contains the DRD4 and DRD2 genes, as well as a
204 few other genes linked to externalizing behavior, such as NCAM1, BDNF, and TPH1 [55], [56],
205 [57]. Limiting the analysis to a smaller number of SNPs allowed us to include all tools except
206 EpiSNP in the comparative analysis (Table 1). EpiSNP encountered execution errors, likely due
207 to the large dataset size. The data contained 446 SNPs with missing values. PLINK (Epistasis
208 and BOOST), MIDESP, and REMMA were used with missing data as is. For other tools, missing
209 SNPs were imputed with the most common non-missing values. Detailed information regarding
210 the genetic and externalizing phenotype data can be found in the ABCD study release notes
211 website.

212 A Bonferroni correction was applied to account for multiple testing, setting the p-value
213 significance threshold at 1.64e-10 for the analysis of 24,699 SNPs. A suggestive significance
214 threshold of 1.64e-9 was also used. For tools that did not provide p-values, such as MIDESP,
215 MDR, and QMDR, we retained the top 0.001% of reported results for MIDESP or the top 500
216 results for MDR and QMDR, based on their respective evaluation metrics (mutual information, t-
217 statistic, and balanced accuracy). While in practice additional steps could be taken to establish
218 the significance of the results reported by these tools, the goal of this study was to evaluate the
219 tools based on their direct output. The interacting SNPs were then mapped onto genes using the
220 UCSC hg19 gene annotation.

221 **Epistasis Tool Selection**

222 We began by conducting an extensive search for epistasis detection tools that could
223 analyze quantitative phenotype data. We focused on publications offering ready-to-use software
224 packages with available documentation. We identified ten tools that were accessible online
225 (Table 1). Since linear regression model is widely used, we evaluated multiple tools based on this

226 model. For other epistasis detection models, we chose the most representative tool for evaluation.
227 After several attempts to install the tools on a personal computer and a high-performance
228 computing (HPC) cluster, we decided to evaluate EpiSNP, Matrix Epistasis, MIDESP, PLINK
229 Epistasis, QMDR, and REMMA. These tools were well documented, simple to install and use,
230 and represented a diverse set of epistasis detection models. PLINK supports multiple epistasis
231 models, and we used its linear regression model for quantitative datasets and its BOOST
232 implementation for case-control datasets. Similarly, we applied QMDR to both quantitative and
233 case-control datasets. We selected BOOST and QMDR for analysis of discretized data, based on
234 previous reviews of epistasis detection algorithms [19]. For simplicity, we refer to QMDR as
235 MDR when applied to binary phenotype data. All selected tools were well optimized and capable
236 of analyzing > 1000 SNPs at a time.

237 **Epistasis Detection Models**

238 *Linear regression* is a linear model used to predict the value of a scalar independent
239 variable based on dependent variable(s). It is used as the underlying model in many epistasis
240 detection tools due to its ability to approximate various types of interactions while being easy to
241 implement and use. The model is similar to the linear regression used in GWAS, except that it
242 includes a term representing SNP-SNP interaction. This allows for a direct interpretation of the
243 results [16]. Both PLINK Epistasis and Matrix Epistasis use this model to estimate interaction
244 coefficients for each SNP pair. The key advantages of this model are its simplicity,
245 interpretability, and fast execution. However, it has limitations, including the assumption of
246 linear relationship between variables and its susceptibility to outliers.

247 *Linear mixed model* (LMM) is an extension of linear regression that allows for both fixed
248 and random effects. LMMs offers significant advantages over traditional linear regression,

249 particularly in genetic studies, as they can account for population stratification and individual
250 relatedness [28]. The increased complexity of LMM, however, results in longer execution times
251 compared to simpler models like linear regression. REMMA employs the linear mixed model in
252 epistasis detection.

253 *General linear model* (GLM) is a broad class of models that includes linear regression,
254 analysis of variance (ANOVA), and analysis of covariance (ANCOVA) [58]. Linear regression is
255 a specific case of general linear models. EpiSNP utilizes GLMs to approximate epistasis, which
256 is similar to, but not identical to, the approach used by PLINK and Matrix Epistasis. The
257 advantages and disadvantages of GLMs are largely similar to those for linear regression.

258 *Mutual Information model* is based on information theory. It quantifies the amount of
259 information obtained about one random variable by observing another. Mutual information (MI)
260 has a wide variety of applications [31]. MIDESP computes the amount of mutual information
261 shared between a subset of most promising SNP pairs and the phenotype. As a feature selection
262 method, MI does not require any prior assumptions and can be used for both categorical and
263 continuous data. However, its main drawback is that it can be computationally expensive to
264 compute.

265 *Multifactor Dimensionality Reduction (MDR)* model is a dimensionality reduction
266 approach for detecting interactions between independent variables influencing a dependent
267 variable [35]. For a pair of biallelic SNPs, MDR considers nine possible genotype combinations:
268 0-0, 0-1, 0-2, 1-0, 1-1, 1-2, 2-0, 2-1, and 2-2 (numbers represent the number of minor alleles).
269 For each of these combinations, a ratio of cases to controls is calculated. Genotypes where cases
270 outnumber controls are labeled as high risk, while the opposite are labeled as low risk. Thus, a
271 pair of SNPs, is represented as a feature with two values: low risk and high risk. A machine

272 learning technique is then applied to identify how useful each feature is for distinguishing
273 between cases and controls. QMDR extends this algorithm to quantitative phenotypes, by
274 comparing mean phenotype values instead of case-control ratios. MDR transforms the data and
275 combines it with a machine learning (ML) classifier to detect epistasis. The strength and
276 weakness of this approach depend on the choice of ML classifier used.

277 **Data Wrangling**

278 The JSON files generated by EpiGEN had to be transformed into appropriate format for
279 each of the selected epistasis detection tools. Several tools (MIDESP, Plink, and REMMA) use
280 standard PLINK input files, while the rest require custom formatting.

281 **Performance Evaluation**

282 All of the selected epistasis detection tools returned lists of interacting SNP pairs.
283 EpiSNP, Matrix Epistasis, PLINK (Epistasis and BOOST), and REMMA provided *p*-values for
284 each SNP pair, while MIDESP provided mutual information values. QMDR provided balanced
285 accuracy values for binary phenotype data, and t-statistic for quantitative phenotype data. We
286 used these values to sort the SNP pairs from most to least significant. *P*-values were sorted in
287 ascending order, while the other metrics were sorted in descending order. Each tool had a
288 mechanism for limiting the number of output SNP pairs. For tools that used *p*-values as the
289 sorting metric, a reporting threshold of 1e-7 was used. MIDESP reported the top 1% of hits by
290 default, and MDR and QMDR reported up to top the 500 hits. Full details regarding the runtime
291 configuration of each tool can be found in supplemental file S2.

292 The sorted lists of hits were then assessed based on three criteria: true positives, the
293 average ranking of the true positives, and the F1 score. The average ranking of the true positives

294 was computed using two approaches: penalized and unpenalized. In the penalized approach, for
295 datasets with more than one interacting SNP pair, any undetected pairs were considered to be at
296 the end of the sorted hit list. In unpenalized approach, undetected pairs were ignored. In datasets
297 with zero true positive, the average true positive ranking was considered not applicable.

298 Results

299 True Positives Detected

300 We examined the number of true positives detected by each tool based on the type of
301 interaction in the EpiGEN simulated data (Figure 1). The true positive percentage (%) refers to
302 the number of successfully detected interacting SNP pairs divided by the total number of truly
303 interacting SNP pairs present in the data. Matrix Epistasis, PLINK Epistasis, and REMMA
304 exhibited nearly identical results, as expected, due to their similar underlying models. These
305 tools excelled at detecting dominant epistasis interactions, identifying 100% of the true positives.

306 MIDESP demonstrated a drastically different performance profile compared to Matrix
307 Epistasis, PLINK Epistasis, and REMMA. It was unable to detect recessive interactions and only
308 identified 33% of the dominant interactions. However, MIDESP performed better at detecting
309 multiplicative and XOR interactions, with detection rates of 41% and 50%, respectively. MDR
310 achieved the best overall performance, with approximately 60% detection. It successfully
311 identified 54% of all multiplicative, 50% of all dominant, 33% of all recessive, and 84% of all
312 XOR interactions. EpiSNP performed best for recessive interactions, with a 66% detection rate,
313 but failed to detect any other interaction types. In contrast, BOOST and QMDR exhibited

generally poor performance across all interaction types, as shown in Figure 1.

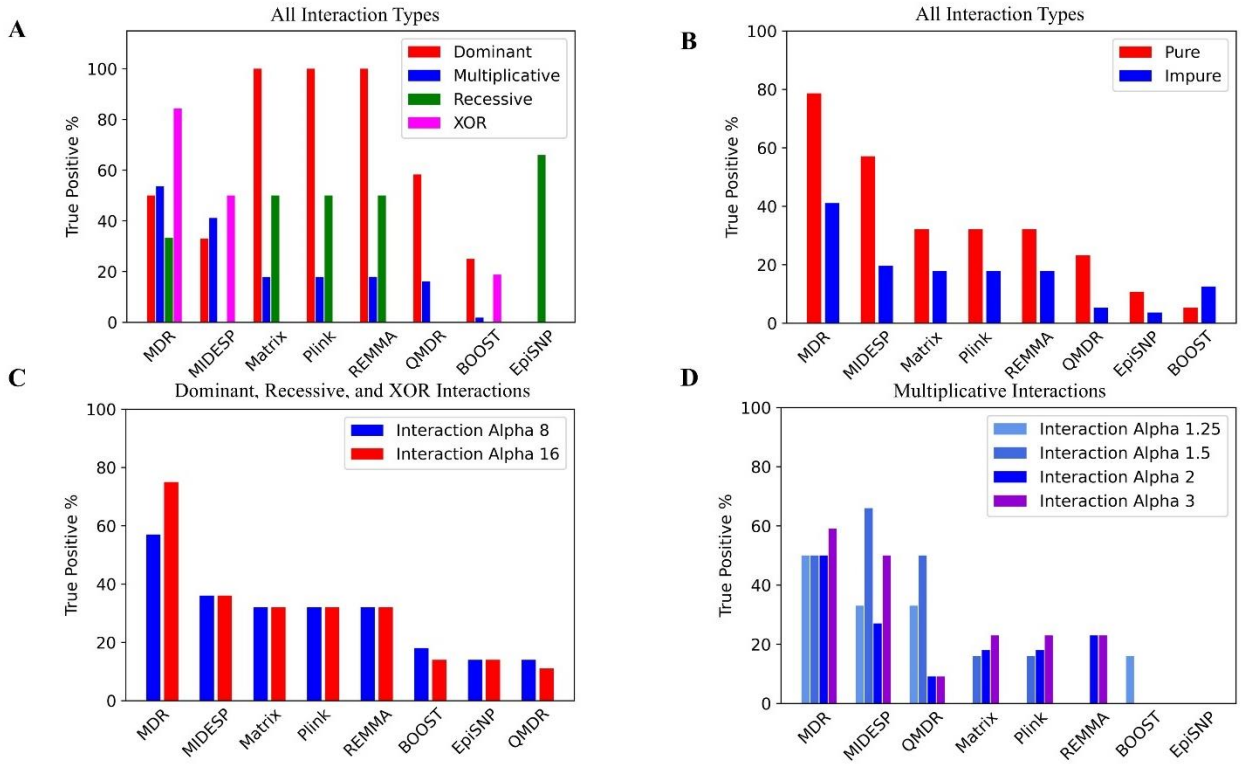

Figure 1: The number of successfully detected interacting SNP pairs (True Positive %) by tools and epistasis interaction types. A. Overall performance of each tool for each *interaction type.* ***B.*** *Pure versus impure datasets.* ***C.*** *Dominant, recessive, and XOR datasets sub-grouped by interaction alpha values: 8 and 16.* ***D.*** *Multiplicative datasets sub-grouped by interaction alpha values: 1.25, 1.5, 2, and 3.*

Average True Positive Ranking

While the number of successful true positive detections is important, such detections are more valuable if they are ranked highly on the sorted list of pairwise SNPs (i.e., the most significant). Otherwise, distinguishing true positives from false positives become challenging. Therefore, for each tool, for each dataset, we calculated the unpenalized average true positive ranking (the average ranking across all interacting pairs within a single dataset) based on the interaction type (Figure 2). Datasets where no true positives were detected were excluded from this analysis. For all interaction types, except recessive, most true positives were ranked toward

the top of the respective sorted lists. The mean and median unpenalized average true positive rankings for each tool and interaction type are summarized in Table 2.

For multiplicative interactions, both MDR and MIDESP show high detection rates and high median average true positive rankings. For dominant interactions, Matrix Epistasis, PLINK Epistasis, and REMMA stand out as the top performers, exhibiting perfect detection rates and average true positive placements. For recessive interactions, EpiSNP is arguably the best tool with the highest detection rate and the best median average true positive placement. For XOR interactions, MDR demonstrates a high detection rate with poor average true positive placement. BOOST on the other hand achieves a high average true positive placement but has a poor detection rate, while MIDESP shows moderate performance in both detection rate and true positive placement.

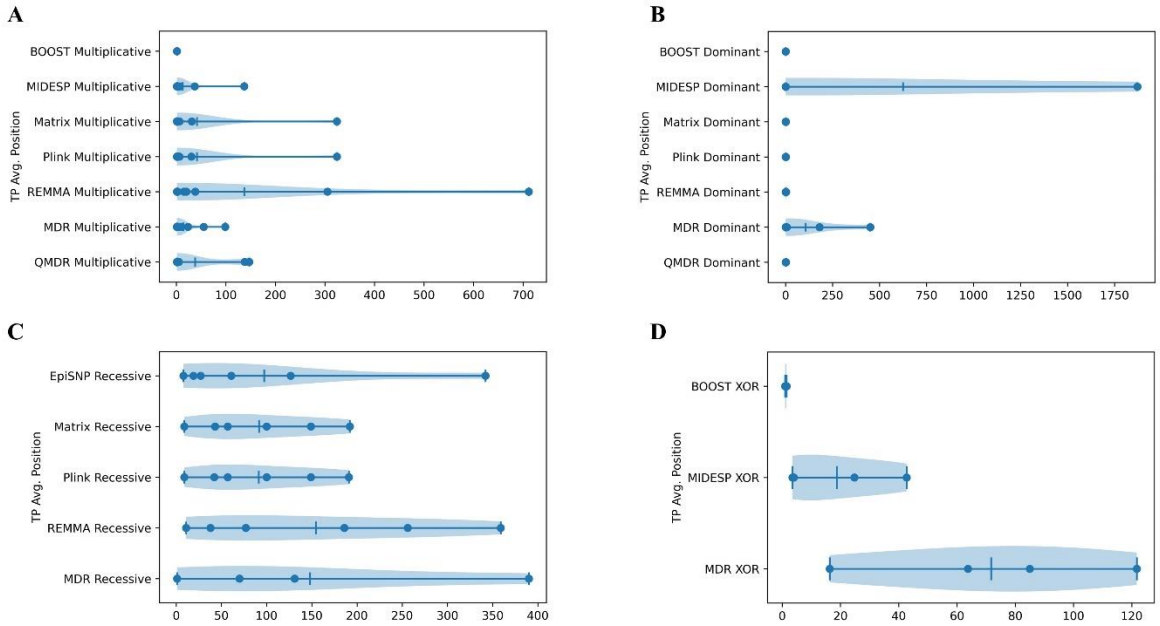

Figure 2: The distribution of unpenalized true positive rankings is shown as violin charts for each epistasis detection tool, categorized by interaction type: A*. Average true positive ranking for multiplicative datasets.* ***B****. Average true positive ranking for dominant datasets.* ***C****. Average true positive ranking for recessive datasets.* ***D****. Average true positive ranking for XOR datasets.*

Table 2: Summary of the mean and median unpenalized average true positive (TP) ranking for each tool and interaction type. *The means and medians are listed in the following order: multiplicative (Mul.), dominant (Dom.), recessive (Rec.), and XOR. If no true positives were identified for a given interaction type, the mean/median are marked as not applicable. Values below 10 are bolded.*

Tool	**Mean TP Position**				**Median TP Position**			
	Mul.	**Dom.**	**Rec.**	**XOR**	**Mul.**	**Dom.**	**Rec.**	**XOR**
MIDESP	12.86	624.5	NA	18.75	**2.5**	**1**	NA	14.4
Matrix Epistasis	41.3	**1.25**	91.67	NA	**3**	**1.25**	78.5	NA
PLINK Epistasis	41.22	**1.25**	91.33	NA	**3**	**1.25**	78.5	NA
PLINK BOOST	**1**	**1**	NA	1.25	**1**	**1**	NA	**1.25**
EpiSNP	NA	NA	97.25	NA	NA	NA	44	NA
MDR	13.72	107	148	71.69	**2.75**	**4**	100.5	74.37
QMDR	37.62	**1.25**	NA	NA	**3**	**1**	NA	NA
REMMA	136.81	**1.625**	154.5	NA	17.5	**1.5**	131,5	NA

F1 Score

To further examine the ratio of true positives, false positives, and false negatives, we computed the F1 scores for each dataset. The mean F1 scores, averaged by interaction type for each tool, are shown in table 3. The means were computed using only datasets where at least one true positive was detected. As indicated by the low F1 values, all tools except for PLINK BOOST generated a large number of false positives (number of false negatives was negligible and can be disregarded). In this regard, PLINK BOOST is the best-performing tool.

Table 3: Summary of mean F1 scores for each tool, for each interaction type. *Only datasets wherein at least one true positive was detected were included in the computation of the mean. Interactions for which no true positives were detected are listed as not applicable.*

Tool	**Mean F1 Score**			
	Mul.	**Dom.**	**Rec.**	**XOR**
MIDESP	0.042	0.0036	NA	0.04
Matrix Epistasis	0.038	0.01	0.002	NA
PLINK Epistasis	0.032	0.0099	0.0019	NA
PLINK BOOST	1	0.66	NA	0.3
EpiSNP	NA	NA	0.0014	NA
MDR	0.0075	0.004	0.004	0.026
QMDR	0.0045	0.0047	NA	NA

REMMA	0.0044	0.005	0.0014	NA

Impact of Purity Status and Interaction Alpha

For most tools and interaction types, the detection rate was higher in pure datasets without individual main effects. The impact of interaction alpha on detection rate varied across tools. The number of interactions detected by each tool, categorized by dataset purity status and interaction alpha value, are shown in Figure 1. A detailed breakdown of each tool's performance by interaction alpha and dataset purity status can be found in supplemental file 3.

Runtime

We briefly evaluated the runtimes of PLINK Epistasis across a range of SNP and sample sizes. PLINK Epistasis was chosen for this evaluation because it is one of the most highly optimized tools included in the study. All analyses were executed on an HPC cluster using one node, one task, and 16 CPUs per task as assigned by the SLURM scheduler [59]. As expected, the runtimes grew approximately following an $O(N^2)$ order with respect to the number of SNPs (N) (Table 4). Runtimes grew linearly, O(M), with respect to the number of individuals (M). Based on these runtimes, analyzing ~1,000,000 SNPs is feasible. In cases where analysis time becomes prohibitive, non-exhaustive algorithms might provide faster runtimes [21].

Table 44: PLINK Epistasis runtimes as measured on HPC (one node, one task, 16 CPUs per task) for simulated datasets with varying numbers of SNPs (1000, 10000, 100000) and individuals (1000, 5000, 10000). *The estimated values are bolded and underlined.*

# of SNPs	# of Individuals	Runtime
1,000	1,000	0.482 seconds
10,000	1,000	6.76 seconds
100,000	1,000	7.75 minutes

10,000	5,000	22.2 seconds
10,000	10,000	43.1 seconds
1,000,000	**10,000**	**129 hours**

ABCD Dataset

Of the six tools used to analyze ABCD data, only four provided p-values: PLINK Epistasis, PLINK Boost, Matrix Epistasis, and REMMA. The suggestive significance cutoff of 1.64e-9 was applied to the results. Using this cutoff, PLINK Epistasis reported 14 suggestively significant results, PLINK BOOST reported 178, Matrix Epistasis reported 26, and REMMA reported 2. For MIDESP, the top 0.001% of results (which contained 180 SNP pairs) was examined. For MDR and QMDR, the top 500 results were reviewed. The interacting SNPs reported by PLINK Epistasis contained rs74991942, which maps onto the DRD4 gene, while the interacting SNPs from PLINK BOOST included rs7131056, which maps onto the DRD2 gene. None of the other reported SNPs from any tool mapped to genes previously associated with externalizing behavior.

The approximate runtimes for each tool are shown in Table 5. These runtimes are provided as examples, not for direct comparison, as different hardware was used for each tool. Furthermore, REMMA can be easily parallelized to run on multiple machines using its built-in functionality, and Matrix Epistasis can also be parallelized, according to the authors [39]. MIDESP can be executed with more threads than we used in this study.

Table 5: Approximate runtimes for epistasis analysis of ABCD dataset consisting of 11,666 individuals and 24,699 SNPs.

Tool	Approximate Runtime	Setup
PLINK Epistasis	5 minutes	High performance cluster (one node, one task, 16 CPUs per task). The CPUs present on HPC are Intel Quad Core E5-2450,
REMMA	120 minutes	

PLINK Boost	180 minutes	E5-2650, E5530 XEON-EMT, E5520 XEON-EMT, and E5430 XEON-EMT. Plink uses up to 15 threads by default. MIDESP uses 8 threads by default. REMMA execution was explicitly subdivided into 16 individual parallel components.
MIDESP	510 minutes	
MDR	480 minutes	Single computer with Intel i7-9700K CPU @ 3.60GHz, 8 cores, up to 8 threads.
Matrix Epistasis	480 minutes	
QMDR	660 minutes	

Discussion

We conducted a survey of the epistasis detection methods for quantitative phenotypes that are suitable for large scale pairwise epistasis scans. Six tools were selected for further evaluation in this study, representing five major models for detecting epistasis: general linear models, linear mixed models, linear regression, mutual information, and multifactor dimensionality reduction. There are some recently developed tools that did not make it into the study, but are worth mentioning such as: Epi-MEIF and QuadKAST [60], [61]. Both tools are well documented and readily accessible online. While this study is not exhaustive, our focus was on providing practical recommendations to the research community, prioritizing usability and optimization over theoretical completeness.

Although the selected tools were highly optimized, processing datasets with millions of SNPs and tens of thousands of individuals remain computationally demanding. Moreover, testing trillions of pairwise interactions can quickly diminish statistical power due to the large number of multiple comparisons that must be corrected. Therefore, a filtering step may be essential to improve efficiency and maintain power. It has been shown that, despite the vast number of potential interacting gene pairs, the actual genetic interaction density is relatively low [62], and not all statistical interactions are biologically meaningful. To address this challenge, an exhaustive search method may be combined with a two-stage approach: first, a subset of genetic

415 variants is selected in a filtering step, and then interactions are only tested among the variants
416 within that subset. The filtering process may rely on statistical tests or biological knowledge. For
417 instance, only genetic variants with significant main effects are subjected to interaction tests [63].
418 Alternatively, individual interactions could be evaluated only among variants showing significant
419 marginal epistasis effects, which refers to the combined pairwise interaction effects between a
420 given variant and all others [64]. The filtering stage can also incorporate biological knowledge,
421 such as pathways linked to diseases, protein-protein interaction networks, expression quantitative
422 trait loci (eQTL) networks, and other regulatory networks. Statistical filtering has the advantage
423 of being discovery-driven and mostly unbiased, whereas biological filtering is hypothesis-driven,
424 making the findings more interpretable. We have successfully used the latter strategy, coupled
425 with a bi-clustering algorithm, to identify interacting genes associated with alcohol use disorders
426 in an admixed population using REMMA [65].

427 In this study, we evaluated the performance of epistasis detection methods using
428 simulated datasets, but we recognize that real-life datasets pose additional challenges. In real-life
429 applications, factors such as covariates, population structure, and family structure must be
430 accounted for [53]. Notably, of all the tools analyzed, only a few allow the inclusion of
431 covariates directly into the epistasis analysis. Moreover, real-life datasets may contain multiple
432 types of interactions simultaneously, whereas our simulation data focused on a specific type of
433 interaction for each dataset.

434 The application of these tools to the ABCD data highlighted their advantages and
435 limitations when handling real-world data. Tools that provide p-values for each SNP pair
436 simplify the process of establishing significance. In contrast, other tools require additional, and
437 sometimes time consuming steps such as permutation testing to establish significance [66]. Real

438 datasets also tend to be much larger than our simulated ones, with many containing millions of
439 SNPs and an increasing number of individuals. Among the tools in this study, PLINK Epistasis
440 and REMMA are the only ones that can be easily applied to large datasets with millions of SNPs
441 and ten thousand individuals. While REMMA is slower than PLINK, it offers the advantage of
442 being highly parallelizable. When analyzing the ABCD data, only PLINK Epistasis and BOOST
443 yielded the results that mapped to genes previously associated with externalizing behavior.
444 Additionally, population stratification, individual relatedness, and other covariates such as sex
445 and age are crucial considerations in real data analysis. REMMA stands out in this regard, as it
446 can directly account for all of these factors. While some other tools can handle covariates, they
447 may not fully account for complex population structures or individual relatedness directly.

448 Our results demonstrate that identifying interacting SNP pairs is challenging due to the
449 large number of false positives reported by each detection method. While true positives are
450 generally found toward the top of the sorted hit lists, this is not always the case. Consequently,
451 further validation, including computational and biological validation, is necessary when
452 interpreting results from epistasis detection analysis. Additionally, it is important to remember
453 that even true statistical epistasis interactions may not always translate into meaningful
454 biological interactions.

455 Our study has demonstrated that existing epistasis detection methods generally provide
456 promising results for quantitative traits. There is however a lack of consistency in the findings
457 across various models, which is not surprising given that the modeling frameworks differ in their
458 respective abilities to detect specific types of epistasis. Based on our analysis, we recommend
459 using a consensus of multiple methods to identify interacting SNP pairs in datasets with
460 quantitative phenotypes [67]. This approach is preferable to relying on a single tool, as we cannot

461 predict which types of epistasis may be present in real data samples *a priori*, and none of the
462 methods evaluated in this study excelled at detecting all interaction types. However, if a single
463 tool is preferred for convenience, PLINK Epistasis and REMMA are the most readily applicable.
464 PLINK is best suited for datasets with low or no individual relatedness, while REMMA is
465 uniquely suited for datasets with complex population structures and high level of individual
466 relatedness. Notably, discretizing a quantitative phenotype (when applicable) followed by
467 applying case-control epistasis detection methods is a viable approach.

468 Key Conclusions and Recommendations:

469 - No single tool excels at detecting all types of epistasis for quantitative phenotypes.
470 - Matrix Epistasis, PLINK Epistasis, and REMMA are best at detecting dominant
471 interactions.
472 - MDR and MIDESP perform best at identifying multiplicative and XOR interactions.
473 - EpiSNP excels at detecting recessive interactions.
474 - PLINK Epistasis and REMMA are well suited for and readily applicable to real-world
475 data.
476 - Combining tools may be the most effective approach, as the interaction types are
477 typically unknown in real datasets.

Funding

This work was supported by the National Institutes of Health (NIH): National Institute on Alcohol Abuse and Alcoholism (NIAAA) T32 AA007456 and National Institute on Drug Abuse (NIDA) DP1 DA054373. NIAAA and NIDA had no further role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development[SM] (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from ABCD Dataset Data Release 5.1.

Data Availability

All of the EpiGEN simulated data and code used to process it can be found in the following GitHub repository: https://github.com/staslist/Epistasis_Review.

Author Biographies

Stanislav Listopad is a postdoc at Scripps Research under the supervision of professor Qian Peng. His research interests lie primarily in the application of software and machine learning engineering to analysis of genomic data.

Qian Peng is a professor of neuroscience at Scripps Research. Her research interests lie primarily in dissecting the complex genomic and epigenomic basis of addiction and related mental disorders through integrative computational approaches.

References

[1] A. Corvin, N. Craddock, and P. F. Sullivan, “Genome-wide association studies: a primer,” *Psychol. Med.*, vol. 40, no. 7, pp. 1063–1077, Jul. 2010, doi: 10.1017/S0033291709991723.

[2] E. Uffelmann *et al.*, “Genome-wide association studies,” *Nat. Rev. Methods Primer*, vol. 1, no. 1, pp. 1–21, Aug. 2021, doi: 10.1038/s43586-021-00056-9.

[3] T. A. Manolio *et al.*, “Finding the missing heritability of complex diseases,” *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009, doi: 10.1038/nature08494.

[4] J. Burch *et al.*, “Wright was right: leveraging old data and new methods to illustrate the critical role of epistasis in genetics and evolution,” *Evolution*, vol. 78, no. 4, pp. 624–634, Apr. 2024, doi: 10.1093/evolut/qpae003.

[5] P. Singhal, S. S. Verma, and M. D. Ritchie, “Gene Interactions in Human Disease Studies—Evidence Is Mounting,” *Annu. Rev. Biomed. Data Sci.*, vol. 6, no. Volume 6, 2023, pp. 377–395, Aug. 2023, doi: 10.1146/annurev-biodatasci-102022-120818.

[6] E. E. Eichler *et al.*, “Missing heritability and strategies for finding the underlying causes of complex disease,” *Nat. Rev. Genet.*, vol. 11, no. 6, pp. 446–450, Jun. 2010, doi: 10.1038/nrg2809.

[7] R. A. Fisher, “XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.,” *Earth Environ. Sci. Trans. R. Soc. Edinb.*, vol. 52, no. 2, pp. 399–433, Jan. 1919, doi: 10.1017/S0080456800012163.

[8] S. A. Kauffman, “Title Pages,” in *The Origins of Order: Self-Organization and Selection in Evolution*, S. A. Kauffman, Ed., Oxford University Press, 1993, p. 0. doi: 10.1093/oso/9780195079517.002.0001.

[9] J. H. Moore, “The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases,” *Hum. Hered.*, vol. 56, no. 1–3, pp. 73–82, Nov. 2003, doi: 10.1159/000073735.

[10] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, “Network medicine: a network-based approach to human disease,” *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011, doi: 10.1038/nrg2918.

[11] J. K. Pickrell, “Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits,” *Am. J. Hum. Genet.*, vol. 94, no. 4, pp. 559–573, Apr. 2014, doi: 10.1016/j.ajhg.2014.03.004.

[12] J. H. Moore, “The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases,” *Hum. Hered.*, vol. 56, no. 1–3, pp. 73–82, Nov. 2003, doi: 10.1159/000073735.

[13] M. T. W. Ebbert, P. G. Ridge, and J. S. K. Kauwe, “Bridging the Gap between Statistical and Biological Epistasis in Alzheimer’s Disease,” *BioMed Res. Int.*, vol. 2015, no. 1, p. 870123, Jan. 2015, doi: 10.1155/2015/870123.

[14] Q. Zhang, J. Liu, H. Liu, L. Ao, Y. Xi, and D. Chen, “Genome-wide epistasis analysis reveals gene–gene interaction network on an intermediate endophenotype P-tau/Aβ42 ratio in ADNI cohort,” *Sci. Rep.*, vol. 14, no. 1, p. 3984, Feb. 2024, doi: 10.1038/s41598-024-54541-8.

[15] M. M. Abd El Hamid, M. Shaheen, Y. M. K. Omar, and M. S. Mabrouk, “Discovering epistasis interactions in Alzheimer’s disease using integrated framework of ensemble learning and multifactor dimensionality reduction (MDR),” *Ain Shams Eng. J.*, vol. 14, no. 7, p. 101986, Jul. 2023, doi: 10.1016/j.asej.2022.101986.

[16] H. J. Cordell, “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans,” *Hum. Mol. Genet.*, vol. 11, no. 20, pp. 2463–2468, Oct. 2002, doi: 10.1093/hmg/11.20.2463.

[17] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, “A survey about methods dedicated to epistasis detection,” *Front. Genet.*, vol. 6, Sep. 2015, doi: 10.3389/fgene.2015.00285.

[18] M. D. Morgan *et al.*, “Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability,” *Nat. Commun.*, vol. 9, no. 1, p. 5271, Dec. 2018, doi: 10.1038/s41467-018-07691-z.

[19] D. Russ *et al.*, “Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models,” *PLOS ONE*, vol. 17, no. 2, p. e0263390, Feb. 2022, doi: 10.1371/journal.pone.0263390.

[20] C. Chatelain, G. Durand, V. Thuillier, and F. Augé, “Performance of epistasis detection methods in semi-simulated GWAS,” *BMC Bioinformatics*, vol. 19, no. 1, p. 231, Jun. 2018, doi: 10.1186/s12859-018-2229-8.

[21] C. Ponte-Fernández, J. González-Domínguez, A. Carvajal-Rodríguez, and M. J. Martín, “Evaluation of Existing Methods for High-Order Epistasis Detection,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 2, pp. 912–926, Mar. 2022, doi: 10.1109/TCBB.2020.3030312.

[22] Y. Wang, G. Liu, M. Feng, and L. Wong, “An empirical comparison of several recent epistatic interaction detection methods,” *Bioinformatics*, vol. 27, no. 21, pp. 2936–2943, Nov. 2011, doi: 10.1093/bioinformatics/btr512.

[23] W. Ren and Z. Liang, “Review on GPU accelerated methods for genome-wide SNP-SNP interactions,” *Mol. Genet. Genomics*, vol. 300, no. 1, p. 10, Dec. 2024, doi: 10.1007/s00438-024-02214-6.

[24] D. B. Blumenthal, L. Viola, M. List, J. Baumbach, P. Tieri, and T. Kacprowski, “EpiGEN: an epistasis simulation pipeline,” *Bioinformatics*, vol. 36, no. 19, pp. 4957–4959, Dec. 2020, doi: 10.1093/bioinformatics/btaa245.

[25] Z. Sha *et al.*, “Distinct network patterns emerge from Cartesian and XOR epistasis models: a comparative network science analysis,” *BioData Min.*, vol. 17, no. 1, p. 61, Dec. 2024, doi: 10.1186/s13040-024-00413-w.

[26] “Epistasis and Its Effects on Phenotype | Learn Science at Scitable.” Accessed: Mar. 22, 2025. [Online]. Available: http://www.nature.com/scitable/topicpage/epistasis-gene-interaction-and-phenotype-effects-460

[27] S. Batista *et al.*, “Computational framework for statistical epistasis supports XOR penetrance function in a living system,” May 09, 2023, *bioRxiv*. doi: 10.1101/2023.05.09.539940.

[28] D. Wang, H. Tang, J.-F. Liu, S. Xu, Q. Zhang, and C. Ning, “Rapid epistatic mixed-model association studies by controlling multiple polygenic effects,” *Bioinformatics*, vol. 36, no. 19, pp. 4833–4837, Dec. 2020, doi: 10.1093/bioinformatics/btaa610.

[29] L. Ma, H. B. Runesha, D. Dvorkin, J. R. Garbe, and Y. Da, “Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies,” *BMC Bioinformatics*, vol. 9, no. 1, p. 315, Jul. 2008, doi: 10.1186/1471-2105-9-315.

[30] “MatrixEpistasis: ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment | Bioinformatics | Oxford Academic.” Accessed: Mar. 22, 2025. [Online]. Available: https://academic.oup.com/bioinformatics/article/34/14/2341/4917352

[31] “MIDESP: Mutual Information-Based Detection of Epistatic SNP Pairs for Qualitative and Quantitative Phenotypes.” Accessed: Mar. 22, 2025. [Online]. Available: https://www.mdpi.com/2079-7737/10/9/921

[32] S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[33] R. De *et al.*, “Identifying gene-gene interactions that are highly associated with Body Mass Index using Quantitative Multifactor Dimensionality Reduction (QMDR),” *BioData Min.*, vol. 8, no. 1, p. 41, Dec. 2015, doi: 10.1186/s13040-015-0074-0.

[34] “BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies: The American Journal of Human Genetics.” Accessed: Mar. 22, 2025. [Online]. Available: https://www.cell.com/ajhg/fulltext/S0002-9297(10)00378-2?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0002929710003782%3Fshowall%3Dtrue

[35] M. D. Ritchie *et al.*, “Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer,” *Am. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–147, Jul. 2001, doi: 10.1086/321276.

[36] H. H. M. Maes *et al.*, “Genetic and Environmental Variation in Continuous Phenotypes in the ABCD Study®,” *Behav. Genet.*, vol. 53, no. 1, pp. 1–24, Feb. 2023, doi: 10.1007/s10519-022-10123-w.

[37] X. Zhang, F. Zou, and W. Wang, “Fastanova: an efficient algorithm for genome-wide association study,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, in KDD ’08. New York, NY, USA: Association for Computing Machinery, Aug. 2008, pp. 821–829. doi: 10.1145/1401890.1401988.

[38] F. Zhang, E. Boerwinkle, and M. Xiong, “Epistasis analysis for quantitative traits by functional regression model,” *Genome Res.*, vol. 24, no. 6, pp. 989–998, Jun. 2014, doi: 10.1101/gr.161760.113.

[39] S. Zhu and G. Fang, “MatrixEpistasis: ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment,” *Bioinformatics*, vol. 34, no. 14, pp. 2341–2348, Jul. 2018, doi: 10.1093/bioinformatics/bty094.

[40] J. R. González *et al.*, “SNPassoc: an R package to perform whole genome association studies,” *Bioinformatics*, vol. 23, no. 5, pp. 654–655, Mar. 2007, doi: 10.1093/bioinformatics/btm025.

[41] V. A. O. Carmelo, L. J. A. Kogelman, M. B. Madsen, and H. N. Kadarmideen, “WISH-R– a fast and efficient tool for construction of epistatic networks for complex traits and diseases,” *BMC Bioinformatics*, vol. 19, no. 1, p. 277, Jul. 2018, doi: 10.1186/s12859-018-2291-2.

[42] R. A. Gibbs *et al.*, “The International HapMap Project,” *Nature*, vol. 426, no. 6968, pp. 789–796, Dec. 2003, doi: 10.1038/nature02168.

[43] N. R. Karcher and D. M. Barch, “The ABCD study: understanding the development of risk for mental and physical health outcomes,” *Neuropsychopharmacology*, vol. 46, no. 1, pp. 131–142, Jan. 2021, doi: 10.1038/s41386-020-0736-6.

[44] H. H. M. Maes *et al.*, “Genetic and Environmental Variation in Continuous Phenotypes in the ABCD Study®,” *Behav. Genet.*, vol. 53, no. 1, pp. 1–24, Feb. 2023, doi: 10.1007/s10519-022-10123-w.

[45] F. Vergunst *et al.*, “Association of Childhood Externalizing, Internalizing, and Comorbid Symptoms With Long-term Economic and Social Outcomes,” *JAMA Netw. Open*, vol. 6, no. 1, p. e2249568, Jan. 2023, doi: 10.1001/jamanetworkopen.2022.49568.

[46] R. Kauten and C. T. Barry, “Externalizing Behavior,” in *Encyclopedia of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford, Eds., Cham: Springer International Publishing, 2020, pp. 1509–1512. doi: 10.1007/978-3-319-24612-3_894.

[47] I. T. Petersen, “Assessing Externalizing Behaviors in School-Aged Children: Implications for School and Community Providers,” *Scanlan Cent. Sch. Ment. Health Pract. Briefs*, vol. 2024, p. https://scsmh.education.uiowa.edu/wp-content/uploads/2024/01/Assessing-Externalizing-Behaviors-in-School-Aged-Children-3.pdf, 2024, doi: 10.17077/rep.006639.

[48] “A gene × gene interaction between DRD2 and DRD4 is associated with conduct disorder and antisocial behavior in males | Behavioral and Brain Functions | Full Text.” Accessed: Feb. 27, 2025. [Online]. Available: https://behavioralandbrainfunctions.biomedcentral.com/articles/10.1186/1744-9081-3-30

[49] S. Hohmann *et al.*, “Evidence for epistasis between the 5-HTTLPR and the dopamine D4 receptor polymorphisms in externalizing behavior among 15-year-olds,” *J. Neural Transm.*, vol. 116, no. 12, pp. 1621–1629, Dec. 2009, doi: 10.1007/s00702-009-0290-1.

[50] E. Koyama, T. Kant, A. Takata, J. L. Kennedy, and C. C. Zai, “Genetics of child aggression, a systematic review,” *Transl. Psychiatry*, vol. 14, no. 1, pp. 1–30, Jun. 2024, doi: 10.1038/s41398-024-02870-7.

[51] “MANUAL FOR THE ASEBA SCHOOL-AGE FORMS & PROFILES.” Accessed: Feb. 27, 2025. [Online]. Available: https://store.aseba.org/MANUAL-FOR-THE-ASEBA-SCHOOL-AGE-FORMS-PROFILES/productinfo/505/

[52] S. Guerrera, D. Menghini, E. Napoli, S. Di Vara, G. Valeri, and S. Vicari, “Assessment of Psychopathological Comorbidities in Children and Adolescents With Autism Spectrum Disorder Using the Child Behavior Checklist,” *Front. Psychiatry*, vol. 10, Jul. 2019, doi: 10.3389/fpsyt.2019.00535.

[53] M. Balvert *et al.*, “Considerations in the search for epistasis,” *Genome Biol.*, vol. 25, no. 1, p. 296, Nov. 2024, doi: 10.1186/s13059-024-03427-z.

[54] J. W. Baurley, C. K. Edlund, C. I. Pardamean, D. V. Conti, and A. W. Bergen, “Smokescreen: a targeted genotyping array for addiction research,” *BMC Genomics*, vol. 17, no. 1, p. 145, Feb. 2016, doi: 10.1186/s12864-016-2495-7.

[55] M. Arcos-Burgos, J. I. Vélez, B. D. Solomon, and M. Muenke, “A common genetic network underlies substance use disorders and disruptive or externalizing disorders,” *Hum. Genet.*, vol. 131, no. 6, pp. 917–929, 2012, doi: 10.1007/s00439-012-1164-4.

[56] C. A. M. Cecil *et al.*, “Neonatal DNA methylation and early-onset conduct problems: A genome-wide, prospective study,” *Dev. Psychopathol.*, vol. 30, no. 2, pp. 383–397, May 2018, doi: 10.1017/S095457941700092X.

[57] C. Xiang *et al.*, “Single nucleotide polymorphisms, variable number tandem repeats and allele influence on serotonergic enzyme modulators for aggressive and suicidal behaviors: A review,” *Pharmacol. Biochem. Behav.*, vol. 180, pp. 74–82, May 2019, doi: 10.1016/j.pbb.2019.03.008.

[58] J. A. Nelder and R. W. M. Wedderburn, “Generalized Linear Models,” *J. R. Stat. Soc. Ser. Gen.*, vol. 135, no. 3, pp. 370–384, 1972, doi: 10.2307/2344614.

[59] A. B. Yoo, M. A. Jette, and M. Grondona, “SLURM: Simple Linux Utility for Resource Management,” in *Job Scheduling Strategies for Parallel Processing*, D. Feitelson, L. Rudolph, and U. Schwiegelshohn, Eds., Berlin, Heidelberg: Springer, 2003, pp. 44–60. doi: 10.1007/10968987_3.

[60] S. Saha, L. Perrin, L. Röder, C. Brun, and L. Spinelli, “Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests,” *Nucleic Acids Res.*, vol. 50, no. 19, p. e114, Oct. 2022, doi: 10.1093/nar/gkac715.

[61] B. Fu, P. Anand, A. Anand, J. Mefford, and S. Sankararaman, “A scalable adaptive quadratic kernel method for interpretable epistasis analysis in complex traits,” *Genome Res.*, vol. 34, no. 9, pp. 1294–1303, Sep. 2024, doi: 10.1101/gr.279140.124.

[62] M. Schuldiner *et al.*, “Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile,” *Cell*, vol. 123, no. 3, pp. 507–519, Nov. 2005, doi: 10.1016/j.cell.2005.08.031.

[63] A. K. Manning *et al.*, “Incorporating biological knowledge in the search for gene × gene interaction in genome-wide association studies,” *BMC Proc.*, vol. 3, no. Suppl 7, p. S81, Dec. 2009.

[64] L. Crawford, P. Zeng, S. Mukherjee, and X. Zhou, “Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits,” *PLOS Genet.*, vol. 13, no. 7, p. e1006869, Jul. 2017, doi: 10.1371/journal.pgen.1006869.

[65] S. Listopad and Q. Peng, “Extensive Genetic Interactions (epistasis) Linked to Alcohol Use Disorder in a High-Risk Population,” Dec. 30, 2024, *medRxiv*. doi: 10.1101/2024.12.29.24319759.

[66] R. De *et al.*, “Identifying gene-gene interactions that are highly associated with Body Mass Index using Quantitative Multifactor Dimensionality Reduction (QMDR),” *BioData Min.*, vol. 8, p. 41, Dec. 2015, doi: 10.1186/s13040-015-0074-0.

[67] S. D’Silva, S. Chakraborty, and B. Kahali, “Concurrent outcomes from multiple approaches of epistasis analysis for human body mass index associated loci provide insights into obesity biology,” *Sci. Rep.*, vol. 12, no. 1, p. 7306, May 2022, doi: 10.1038/s41598-022-11270-0.