High-order interaction analysis in genome-wide association studies using multifactor dimensionality reduction

Sungyoung Lee¹, Joon Yoon¹, Seungyeoun Lee², and Taesung Park¹,³,§

¹Interdisciplinary Program in Bioinformatics, Seoul National University
³Department of Applied Mathematics, Sejong University
³Department of Statistics, Seoul National University, Seoul, Korea. 151-742.

§tspark@stats.snu.ac.kr

Abstract: Gene-gene interaction (GGI) plays an important role in the causation of complex diseases, and its importance has now been well recognized through the findings of many successful genome-wide association studies (GWAS). Although many statistical methods have been introduced to address GGI analysis in GWAS, these methods have mainly focused on two-way interactions, rather than on high-order interactions. In addition, rapid advancement of biotechnology has significantly increased the number of genetic variants that are detectable, which makes an exhaustive approach unfeasible. In order to overcome the computational challenge of high-order GGI analysis using statistical approach, we develop a novel and efficient strategy called Hi-Mise; a high-order interaction analysis using the Multifactor Dimensionality Reduction (MDR) method with Interaction Set Expansion, for simultaneous identification of high-order interactions. Hi-Mise consists of second-order interaction scanning step, interaction seed initialization step, and interaction set expansion step. These steps have been computationally optimized for detection of high-order interactions. Through simulation studies using real GWAS data, Hi-Mise was shown to be capable of detecting high-order interactions with high testing balance accuracies (BAs). In addition, the application of real GWAS data showed that Hi-Mise could successfully identify multiple high-order interactions simultaneously for cases with and without marginal effects.

Keywords: Gene-gene interaction, genome-wide association study, high-order interaction, graphical processing unit

1 Introduction

Epistasis, or genetic interaction, has been recognized to play an important role in many organisms such as plants and animals, including in humans [1, 2]. Importance of epistasis strongly supports a hypothesis that a complex biological network involving multiple genes, and interactions among those genes, is responsible for many complex traits. In this respect, the identification of gene-gene interactions (GGIs) is essential to understanding the underlying biological mechanisms of complex traits.

In order to identify GGIs, researchers have proposed various methods. Traditionally, GGI studies in an early stage were conducted with logistic regression, using a population-based case-control design. Despite the many successes of logistic regression, the existence of linkage disequilibrium (LD) is a major weakness of this method, because it causes the multicollinearity of single nucleotide variants (SNVs). In contrast to logistic regression, Multifactor Dimensionality Reduction (MDR), proposed by Ritchie et al., is a nonparametric and model-free method for exploring GGIs that is not affected by LD [3]. The usefulness of the MDR method has been demonstrated through the successful discovery of many novel GGIs [4-6], and there have since been various extensions of MDR [7-10].

Analysis of GGIs for GWAS has suffered from computational burdens that were due to large numbers of genetic variants. For example, when the number of variants is 500K, the number of two-way interactions is approximately $10^{13}$. Thus, many GGI methods have employed a two-stage approach to overcome this issue. The first stage consists of the selection of variants with some marginal effects and
the second stage consists of the performance of GGI analysis for the selected variants. In addition to two-stage approaches, a number of alternative methods have been proposed. Wan et al. introduced a computationally efficient method using Boolean operation-based logistic regression framework, which is practical to exhaustively investigate two-way GGIs in large-scale datasets [11]. Although their method has shown impressive performance in terms of computation and statistical power, investigation of high-order interactions with this method is impossible. The incredibly high number of exhaustive combinations over two-way interaction means that the application of an exhaustive approach for high-order interactions is unfeasible. Thus, we previously introduced a toolset called cuGWAM [compute unified device architecture (CUDA)-based genome-wide association MDR], using a graphical processing unit (GPU) to accelerate the analysis performance of MDR [12]. Although cuGWAM performs well in the discovery of two-way interactions, it does not seem to perform efficiently for high-order interactions owing to memory limitations. Consequently, we needed to develop an alternative method to identify high-order interactions in GWAS data.

In recent years, a number of reports have highlighted an importance of GGI for the missing heritability problem for complex traits [13, 14]. In addition, it was shown that an interaction across multiple genes occurs not only in a pair-wise manner, but also via complex network of genes [15, 16]. In this respect, the identification of high-order genetic interaction could be a key process in solving the problem of missing heritability. Unfortunately, only a few statistical methods are currently available for high-order GGI identification. Gao et al. extended the forward LASSO analysis for high-order interactions, and showed that their approach was capable of identifying high-order interactions within a small number of datasets [17]. Nevertheless, since the method is essentially based on LASSO, the number of variants that can be handled is limited due to poor memory availability. Thus, in large-scale data sets, additional manipulation of the dataset is required in order to reduce the number of variants. Another notable method was introduced by Ming et al. and was based on U-statistics with a forward selection algorithm, which can be applied for the discovery of high-order interactions [18]. However, this method has some limitations in that it can only be applied to quantitative traits and can detect only a single interaction per variant. It should be noted that it has been frequently reported that many meaningful and productive interactions share some hub variants or genes.

In addition to the statistical techniques, there are a number of extended statistical methods available for identification of high-order interaction that incorporate known biological information to reduce the search space. Oh et al. proposed a method to identify high-order GGIs using gene-based MDR analysis [8]. Although the gene-based MDR method does enable the identification of high-order interactions, it still suffers from heavy computational burden.

In order to overcome the current computational challenges in identifying high-order GGIs, we developed a novel approach called Hi-Mise (High-order interaction discovery method using MDR with interaction set expansion) for simultaneous identification of high-order GGIs. Hi-Mise consists of a second-order interaction scanning step, an interaction seed initialization step, and an interaction set expansion step. These steps have been computationally optimized for detection of high-order GGIs and Hi-Mise is designed to identify multiple sets of SNVs based on exhaustively investigated second-order interactions, through integrating the set of these interactions and of other SNVs into the dataset.

Through simulation studies using real GWAS data, Hi-Mise was shown to be capable of reliably detecting high-order interactions. Unlike other existing methods, Hi-Mise outperformed the comparable toolsets that could be applied to GWAS datasets, while analyzing an entire dataset with no strategy for reducing the number of SNVs. Furthermore, Hi-Mise has a further potential to include other, existing two-way interaction analysis methods. In the application of Hi-Mise to a real GWAS dataset, we also found that the results produced included substantial overlap with the previous findings. This real data application showed that Hi-Mise could successfully identify multiple high-order interactions simultaneously in cases with or without marginal effects.

2 Method and Materials

2.1. Procedures for the Hi-Mise method

In this procedure we considered the identification of high-order interaction from GWAS data, which consists of a large number of SNVs and thousands of samples, and the binary phenotype of interest. Here,
the goal of our proposed approach was the simultaneous identification of up to \( k^{th} \)-orders of multiple interactions. Since there are extremely high numbers of possible \( k^{th} \)-orders of combinations, as well as its subsequent lower order combinations, \( 2^{nd}, ..., (k-1)^{th} \), an exhaustive investigation of all possible \( k^{th} \)-order interactions was unfeasible. For these reasons, we designed our approach in a computationally feasible manner, as described in Fig. 1.

**Second-order scanning step.** The multiple testing problem is one of the major limitations of the GGI detection methods. In order to avoid the multiple testing issue, our method first enumerated all possible second-order SNV combinations: \( G = \{g_1, g_2, ..., g_{p-1}, g_p\} \) and then performed MDR analysis. Here, \( p \) denotes the number of SNVs in the dataset. For each combination of SNVs in \( G \), an evaluation measure, \( t \), is computed. Many existing measures such as accuracy or balanced accuracy can be used as this evaluative measurement, \( t \). Finally, the top second-order SNV combinations, \( q \), with the best evaluation measures, \( t(1), ..., t(q) \), are extracted. With this strategy, it is possible to select top \(q\) from \( pC_2 \) possible two-way combinations. Here, a value for \( q \) (i.e. 1K or 10K) should be determined by the user, depending on the number of SNVs in the dataset.

**Interaction seed initialization step.** After the second-order scanning step, a set of initial seeds: \( S(1) = \{V(1), U(1)\} \) is constructed. Here, \( V(1) = \{v_1, ..., v_q\} \) corresponded to the top \( q \) interactions \( (g(1), ..., g(q)) \) in which \( g(j) = \{g_{j1}, g_{j2}\} \) had the evaluation measure \( t(j) \), and \( U(1) = \{u_1, ..., u_r\} \) corresponded to a set of SNV nodes consisting of \( r \) SNVs, which are not involved in the top \( q \) interactions. Here the value of \( r \) depended on the value of \( q \). For example, from a dataset with 100 SNVs, if we picked the top 10 interactions comprised of 15 unique SNVs, the value of \( r \) would be 85.

**Interaction set expansion step.** Based on the initial seed set, \( S(1) \), an iterative expansion of the interaction set is performed. In each \( l^{th} \) iteration, the set of interactions is updated from \( S(l) \) to \( S(l+1) \), only if the interactions have at least one edge. Here, the edge is defined as a relationship which connects two nodes if an evaluation measure increases over a certain amount, \( \tau \), by combining them. The threshold of integration for \( \tau \) should be specified by the user. This condition is equivalent to \( M(\{v_l, v_j\}) > max(M(v_l), M(v_j)) + \tau \) or \( M(\{v_l, u_j\}) > max(M(v_l), M(u_j)) + \tau \), where \( M(\{v_l, v_j\}) \) and \( M(\{v_l, u_j\}) \) represent evaluation measures for \( \tau \) with corresponding SNVs for \( v_l, v_j \in V(I) \) and \( u_j \in U(I) \), respectively. Through the following merge, addition, and removal processes, the interaction set is repeatedly updated. First, if two interaction nodes connected with an edge shows an improvement of evaluation measure over \( \tau \), then they are merged into a new node to encompass higher order of interaction (merge). Second, if there is not an interaction node but an SNV node that improves an evaluation measure of an interaction over \( \tau \), a new node is created by combining the interaction node and the SNV node (addition). For example, we let \( M(v_l) = 0.65 \), \( M(v_j) = 0.54 \) and \( M(\{v_l, v_j\}) = 0.68 \), the merge process takes place if \( \tau < 0.03 \). Otherwise, all SNVs in \( U(I) \) would be investigated to find the best \( u_j \) that satisfies \( M(\{v_l, u_j\}) > M(v_l) + \tau \), and the ‘addition’ process takes place if such \( u_j \) exists. Finally, if no such \( u_j \) exists, all SNVs within \( v_l \) would be investigated with the same purpose and the ‘removal’ process takes place. This expansion step allows \( V(I) \) to include other SNVs which are not in the significant interactions, so that the updated interaction set has more diversity. Finally, if the above conditions are not met, an SNV comprising the interaction node would be removed (Removal). For each iteration, the set of interaction nodes \( V(I) \) is stored. The interaction set expansion step is repeated until no change in the interaction set can be made.

**Result export step.** If the interaction set expansion step stops after \( L \) iterations, the final set \( S(L) = \{V(L), U(L)\} \) is made. As a final summary output, all interactions within the stored interactions for each iteration \( V(2), ..., V(L) \) are exported with their evaluation measures.

Since the first step of our approach involved a massive computational burden, we adopted a technique from our previous work using a GPU system [12] into the second-order scanning step. All other steps were implemented for execution using a CPU system.

### 2.2. Real data analysis

We applied Hi-Mise to the analysis of the Wellcome Trust Case Control Consortium (WTCCC) dataset with bipolar disorder (BD) as a phenotype of interest. The WTCCC study was a large-scale GWA study that focused on seven complex diseases: BD,
cardiovascular disease, hypertension, rheumatoid arthritis, Crohn’s disease, and type 1 and type 2 diabetes [19]. Roughly 2,000 samples for each disease were included in the study, along with 3,000 shared common controls, all of European ancestry. The shared controls consisted of two cohorts: a 1958 Birth Cohort (58C) and UK official blood service samples (NBS). All of the individuals were genotyped using Affymetrix GeneChip 500K arrays. We used the genotype dataset generated by the genotype calling algorithm, CHIAMO.

We first conducted a quality control (QC) process for the genotype dataset with the following steps: (1) After the removal of SNVs with a Hardy-Weinberg Equilibrium test p-value lower than $5.7 \times 10^{-7}$ in the control samples, (2) SNVs with a genotypic association test p-value lower than $5.7 \times 10^{-7}$ between the 58C and NBS cohorts were further excluded. (3) Finally, SNVs with minor allele frequency (MAF) lower than 5% and genotyping rate lower than 95% were excluded from the analysis. After this quality control step, the dataset was imputed using fastPHASE software to generate the final dataset. The final dataset used in our analysis contained 354,022 SNVs and 4,806 samples consisting of 1,868 BD and 2,938 control samples.

### 2.3. Simulation study using real dataset

At its core, Hi-Mise is an extension of MDR for the identification of high-order GGI s. In this respect, we conducted a simulation study to validate the performance of Hi-Mise compared to the original MDR. In this simulation study, we focused on two aspects: whether Hi-Mise provided a better evaluation measure compared to the original MDR method and whether Hi-Mise eliminated substantial overlap among the interaction combinations that had the best evaluation measures. In this study, we varied the sample sizes using 100, 200, 500 and 1,000 samples. For each sample size, we generated 10,000 datasets using real GWAS data, described in the above section. Each simulation dataset consisted of 100 SNVs which were randomly chosen from the real, published dataset. Although the number of SNVs we simulated was much less compared to a real GWAS dataset, the number of simulated SNVs was restricted to being relatively small due to the computational limitation of exhaustive MDR. For each set in a simulation, we applied the proposed Hi-Mise, original MDR, and cuGWAM tools, along with the unmodified phenotypes from the real GWAS dataset.

### 3 Results

We implemented our approach using a program called Hi-Mise, which was written in two programming languages: R (http://www.r-project.org/) and C/C++. Hi-Mise can analyze large-scale datasets, and can handle the genetic datasets in various formats, including in a Binary PED file of PLINK or variant calling format (http://www.1000genomes.org/). Hi-Mise is currently freely available from our website (http://bibs.snu.ac.kr/software/himise/). Both the simulation and the real GWAS dataset analyses were conducted with GPU-based systems consisting of a single Intel i7 950 processor, 32 GB of RAM, and three NVIDIA GTX 480 graphics cards.

#### 3.1. Computing performance of the analysis

Our proposed approach solved the computational burden experienced using other techniques by hierarchically integrating second-order interactions. In order to assess the gain in computational performance, we conducted an evaluation of a series of comparisons focusing on the computing times of three toolsets: Hi-Mise, cuGWAM, and the original MDR toolset. The latter two toolsets examined all possible combinations with a given order of interaction for the comparisons. As shown in Table 1 by the 5-fold cross-validation, Hi-Mise showed remarkably faster analysis speed than the original MDR toolset in all orders of interactions. The execution time of Hi-Mise was even shorter than the GPU-based cuGWAM toolset. It should be noted that cuGWAM uses the same algorithm as MDR, but is implemented in a GPU system. cuGWAM showed significantly superior performance to the original MDR [20]. In addition, the computational efficiency of Hi-Mise became remarkably higher than the other methods as the number of samples increased. Here, the computational efficiency was defined as the ratio of computing time of Hi-Mise to that of the original MDR or cuGWAM.

#### 3.2. Comparison of testing balance accuracy

An important feature of Hi-Mise is the ability to accurately detect high-order interactions without exhaustive investigation. In order to assess the detection performance of Hi-Mise, we compared the balanced accuracies (BAs) from testing dataset, between Hi-Mise and exhaustive MDR for all possible orders of interaction. It should be noted that
the result from the GPU-based MDR was not considered in the comparison because the results of original MDR and GPU-based MDR are essentially identical. In the comparison, we mainly focused on both the accuracy in the testing dataset of identified interactions and the identification of interactions without marginal significance. Hence, as the first step, the range of testing BAs of top \(k\) combinations from Hi-Mise and the original MDR were compared. Fig. 2 summarizes the results from 4th order interaction analyses. As shown in Fig. 2A and 2B, the top \(k\) testing-BAs of Hi-Mise (0.6 ~ 0.8) were similar to those for exhaustive MDR (0.47 ~ 0.8) for the same orders of interaction. However, the range and average of testing-BAs from Hi-Mise were considerably higher than those from exhaustive MDR. This implies that the proposed method is capable of identifying interactions with greater BA than the original MDR method.

Since the majority of top \(k\) high-order interactions from exhaustive MDR analyses included SNVs with large marginal effects, an interaction that did not contain marginally significant SNVs tended to be excluded from top \(k\) selection in the original MDR. In contrast to exhaustive MDR, the interaction set expansion step of Hi-Mise prevented the SNVs with marginal effects from having a dominating influence on the top \(k\) results. Thus, the presence of SNVs with large marginal effects had less influence on Hi-Mise. Fig. 2C and 2D illustrate the effects of SNVs with large marginal effects. Here, we defined the influence of these marginal effects as the frequency of an SNV among the selected interactions. While the largest frequency of SNV in terms of nodes (i.e. hub SNVs) by Hi-Mise was 135, the highest frequency by exhaustive MDR was 197. This difference between the two highest frequencies indicates that Hi-Mise was less affected by hub SNVs with large marginal effects.

Furthermore, our approach simultaneously discovered multiple high-order interactions with different orders in a single run, whereas exhaustive MDR was only capable of identifying the interactions with a given order for each run.

3.3. Real data analysis

Using the WTCCC dataset after quality control, we performed high-order GGI analysis, up to eighth order, using Hi-Mise. At first, we performed the second-order interaction scanning step and identified the top 10,000 two-way interactions with the highest testing BA. Then we generated initial interaction seeds using these top 10,000 two-way interactions and performed the interaction set expansion step. The second-order scanning step took approximately 1.5 days for exhaustive two-way interaction analysis, and the interaction set expansion step took 3 hours with three iterations. The top-ranked interactions from third order to eighth order are listed in Table 2, except for those of fifth order, due to unavailability of this data. As shown in Table 2, the discovered high-order GGIs shared certain common SNVs. However, the link between the shared SNVs and the GGIs was not statistically significant in the single variant logistic regression analysis (\(p=0.005\), \(p=0.859\), \(p=0.008\), \(p=0.026\), \(p=0.005\)). Moreover, the identified high-order interactions showed distinctive patterns among different orders, indicating a possibility of different interaction mechanisms that depend on the orders of interaction.

In order to interpret the identified high-order interactions, we annotated the identified GGIs using several online references, including a GWAS catalog [21], Metamoods [22], and Regulome database [23]. The results of annotation showed that a considerable number of high-order interactions included SNVs which have been evidenced to be associated with BD, or other mood related diseases such as schizophrenia [21-24].

4 Conclusions

Inspired by the important role of epistasis, many novel methods for an identification of GGIs have been proposed. However, only a handful of these methods are available for use in the detection of high-order interactions, especially for binary traits. This shortage of high-order GGI methods is mainly attributable to an enormous number of possible combinations of high-order GGIs in the GWAS dataset. Although MDR has made notable advancements in GGI discovery, the application of MDR to high-order interactions has been impeded by the exponentially growing number of possible combinations [25]. In this paper, we proposed a novel method for the detection of high-order GGIs. It acts in a computationally efficient manner, using a second-order interaction scanning step, interaction seed initialization step, and interaction set expansion step. Unlike the previous methods for high-order interaction analysis, our unified Hi-Mise approach is capable of simultaneously identifying various orders of high-order interactions. In addition, the Hi-Mise toolset provides results concerning the interactions of
multiple orders in a single run. This simultaneous identification can remarkably reduce the computational burden of the analysis. Moreover, the proposed method provides a strategy to avoid the situation where SNVs with large marginal effects have the dominating influence in the results. The results of the simulation using a real GWAS dataset successfully demonstrated the advantage of our proposed approach. In addition to the high accuracy of the high-order interaction detection, our method also showed substantial improvements to computational complexity, even when compared to GPU-based MDR analysis. Because of the massive computational burden, previous methods for high-order GGI identification had employed various strategies to reduce the search spaces, such as gene-based approaches or variant filtering approaches. However, true high-order GGIs could be eliminated by these strategies [26]. In this respect, the proposed method successfully identified high-order GGIs using a GWAS dataset without biological knowledge or variant filtering. Furthermore, our method also showed considerably higher testing BAs in the test datasets than the testing BAs obtained for the original MDR method, demonstrating the usefulness and efficacy of our method.

Hi-Mise method has shown several advantages in both computing time and in the capacity to identify the interactions efficiently. For the dataset that we have analyzed in this paper, the number of possible combinations for exhaustive 8-way MDR would be approximately $6.12 \times 10^{39}$. Thus, the computational time of exhaustive 8-way MDR is literally immeasurable. On the other hand, the analysis time of Hi-Mise was less than 42 h, and this process encompasses the analysis of eighth order interactions and all subsequent lower orders from seventh to second.

The annotation of the results of Hi-Mise using real dataset analysis showed that the identified SNVs were associated with BD and other mood-related diseases. In addition, Hi-Mise also identified the hub SNVs across multiple orders from the results produced. Interestingly, some of these hub SNVs had not been previously reported for BD, and their marginal effects were also moderate. The hub SNVs detected here by our application may suggest that our method is capable of finding novel hub SNVs without marginal effects. For the hub SNVs, analysis using the real dataset up to eighth order yielded distinctive patterns across different orders of interactions. For example, the third and fourth orders of interactions shared hub SNVs, while the fifth and sixth orders shared different hub SNVs, and finally the eighth order interactions did not share any hub SNVs. These results are quite different from those that have been produced by traditional GGI analyses where marginal SNVs consistently influenced the top k results for all orders of interaction.

Moreover, our method can be extended to other GGI detection approaches. Currently, our method does not provide statistical significance such as p-values based on permutation due to the computational complexity involved. In future work, we plan to extend and improve Hi-Mise for a wider range of applications; these improvements involve the inclusion of other parametric methods as well.

In conclusion, we have successfully demonstrated that our proposed method has the following advantages: firstly, it can detect high-order interactions in a GWAS dataset in an adequate computation time. Secondly, for our method we have reported simultaneous multi-order interaction detection in a single run. Thirdly, the results from our method have shown high accuracy of the technique. Finally, our method is capable of identifying hub SNVs that do not have marginal effects. We expect that our method would be useful to researchers in practical applications for the identification of high-order GGIs in large-scale datasets.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP) (No. 2012R1A3A2026438).

References:


Fig. 1. Flow of high-order interaction discovery using Hi-Mise.

**Second-order scanning**

- Compute exhaustive second-order interactions
- Extract top $q$ interactions with evaluation measure $t(1), \ldots, t(q)$
- Set an initial interaction set $S(1) = \{V(1), U(1)\}$ with $q$ interactions

**Interaction set expansion**

- $V(t)$, $U(t)$
- SNVs outside interaction
- Repeat until $V(t) = V(t+1)$
- Find possible transforms
- Merge
- Add
- Remove
- Update $V(t)$ and $U(t)$
- Final interaction set $S(k) = \{V(k), U(k)\}$ after k iterations
- Report identified interactions in $V(2), \ldots, V(k)$

Fig. 2. Results of the simulation. All figures were drawn based on the top 500 results from the methods used. (A) Histogram of prediction accuracy for test data using Hi-Mise. (B) Histogram of prediction accuracy of test data using the exhaustive method. (C) Histogram of the degree of single nucleotide variants (SNVs) using Hi-Mise. (D) Histogram of the degree of SNVs using the exhaustive method. Here the degree of SNVs is the frequency of their observation in the discovered GGIs.
Table 1. Execution time of $k^{th}$-order interaction analysis.

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Order</th>
<th>MDR</th>
<th>cuGWAM</th>
<th>Hi-Mise</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3</td>
<td>1.6</td>
<td>1.4</td>
<td>1.6 (1x)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>55</td>
<td>6.48</td>
<td>1.8 (30.6x)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1250</td>
<td>248.538</td>
<td>4.5 (277.8x)</td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>2.5</td>
<td>1.42</td>
<td>1.6 (1.6x)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>78</td>
<td>8.05</td>
<td>2 (39x)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1944</td>
<td>317.65</td>
<td>6.5 (299.1x)</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>4.9</td>
<td>1.45</td>
<td>1.6 (3.1x)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>154</td>
<td>13.01</td>
<td>2.2 (70x)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3826</td>
<td>532.35</td>
<td>11 (347.8x)</td>
</tr>
<tr>
<td>1,000</td>
<td>3</td>
<td>8.8</td>
<td>1.48</td>
<td>1.7 (5.2x)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>273</td>
<td>20.75</td>
<td>3.4 (80.3x)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7261</td>
<td>885.98</td>
<td>17.5 (414.9x)</td>
</tr>
</tbody>
</table>

Unit: second

* Multifactor dimensionality reduction (MDR) and cuGWAM were performed with exhaustive investigation.
* Hi-Mise execution time includes an identification of subsequent orders of interactions.
* Parenthesized numbers of Hi-Mise are the fold acceleration compared to MDR.

Table 2. Results of the Wellcome Trust Case Control Consortium (WTCCC) dataset with bipolar disorder (BD) phenotype of interest, analyzed using Hi-Mise.

<table>
<thead>
<tr>
<th>Combination</th>
<th>Order</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMP3, LOC102724210, RPL23AP82 (rs514636,rs12642242,rs9616985)</td>
<td>3</td>
<td>0.62</td>
</tr>
<tr>
<td>LAMP3, LOC102724210, ACR (rs6414498,rs12642242,rs3810648)</td>
<td>3</td>
<td>0.62</td>
</tr>
<tr>
<td>CROT, RPL23AP82, SHANK3 (rs2051950,rs9616985,rs739365)</td>
<td>3</td>
<td>0.61</td>
</tr>
<tr>
<td>LAMP3, LOC102724210, FAM184A, RPL23AP82 rs6414498,rs12642242,rs12213597,rs9616985</td>
<td>4</td>
<td>0.62</td>
</tr>
<tr>
<td>NUF2, PLB1, n/a, n/a, THSD7A (rs17363152,rs6748157,rs12637543,rs6838310,rs2789034,rs1859226)</td>
<td>6</td>
<td>0.65</td>
</tr>
<tr>
<td>NUF2, TPO, BSN, n/a, n/a, n/a, THSD7A (rs17363152,rs10204515,rs9858542,rs1304118,rs899628,rs6838310,rs1859226)</td>
<td>7</td>
<td>0.69</td>
</tr>
<tr>
<td>n/a, TPO, n/a, AC007392.3, BSN, n/a, n/a (rs12117214,rs17097137,rs10203313,rs9309393,rs9858542,rs6838310,rs7663402,rs1885452)</td>
<td>8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

* Single nucleotide variants (SNVs) in bold text indicate the existence of known associations
* SNVs in underlined text indicate the existence of linkage disequilibrium (LD) for SNVs with known associations.
* n/a: Gene information for the SNV was not available.