A comparative Genome-Wide Association Interaction Study using BOOST and MB-MDR algorithms on Ankylosing Spondylitis

Kyrylo Bessonov [1][2], Elena Gusareva [1][2], Kristel Van Steen [1][2]
[1] Systems and Modeling Unit, Montefiore Institute, University of Liege, 4000 Liege, Belgium; [2] Bioinformatics and Modeling, GIGA-R, University of Liege, 4000 Liege, Belgium

Introduction

Genome-Wide Association (GWA) studies have gained popularity after the completion of the Human Genome Project and advancement of high-throughput technologies. These studies aim to scan thousands of genomic variations (e.g., SNPs) for their association to phenotypic variables (i.e., traits), such as disease related phenotypes, with the hope of extracting biologically and clinically relevant information. Understanding of genetic, environmental as well as other components of the disease brings the key insights into disease pathophysiology and approaches us closer to the ultimate goal - personalized medicine.

In this work we rely on a minimal GWA protocol for genome-wide epistasis detection using SNPs, as developed in our lab [6][9]. Using the advanced non-parametric Model-Based Multifactor Dimensionality Reduction (MB-MDR) method [1] and Booelan Operation-Based Screening and Testing (BOOST) algorithms, we investigate the effect of exhaustive (BOOST) and non-exhaustive (MB-MDR) marker processing strategies on LD effects, as well as different adjustment schemes for lower-order effects (i.e. epistasis).

Our approach was tested on Ankylosing Spondylitis (AS) data as provided by the WTCCC2 consortium [1]. AS is a long-term / chronic disease characterized by inflammation of the joints between the spinal bones. Non-steroidal anti-inflammatory drugs calming down the immune system inflammatory responses are used as a treatment but there is no permanent cure for AS. The disease has also a strong environmental component and affects 3.5 - 15 per 1,000 people in USA [5].

Methods

The AS SNP data were obtained from the WTCCC2 and a subset consisting of 487,780 SNPs and assigned to 1788 cases and 4799 controls was obtained according to SNP and sample lists given provided in [2]. Thus our input dataset was exactly the same as the one used in the reference study [2]. The overall workflow is shown in Fig. 1 consisting of various methods assessing the effect of Linkage Disequilibrium (LD), association selection and lower-order adjustment schemes in MB-MDR.

Data extraction and LD pruning

To extract the subset from raw data, SNP data extraction was done with PLINK. To avoid an abundance of redundant SNP-SNP interactions (caused by LD between SNPs) we implemented LD pruning strategy via SVS 7.6. This algorithm correlates correlation threshold of 0.75 and window size of 50 bp with 1 bp increment.

Data filtering using Filter 2.0

The search space was reduced to optimize chances of finding truly biologically relevant SNPs (i.e. pre-filtering). This was done using Filter 2.0 [10] adopting two strategies requiring: a) a minimum of 3 data sources [e.g. KEGG, BioGrid, MINT] supporting given SNP-SNP interaction (implication index of 3); b) candidate gene list related to AS pathology and associated pathways including literature reported markers such as HLA-B, IL23R, ERAPI and KIF218 [2].

% overlap between workflows

The final results represented as a list of significant SNPs with corresponding statistics were compared across workflows (Fig.1). The maximal %overlap value between final results (maximum[# of common SNPs pairs/# of total SNPs pairs work flow 1] , # of common SNPs pairs/# of total SNPs pairs workflow 2]) for a selection of workflows are partially reported in Tables 1 and 2.

Ranks calculation

To compare the variability between the outcomes of different workflows, ranks (i.e. positions) were calculated on all outputted SNP pairs across workflows regardless of statistical significance.

Euclidean distance between workflows was found using an input vector of 36 ranks of common SNP pairs from top 2000 in the final MB-MDR results list. The resulting dendrogram (hierarchical tree) shows the impact of different choices (LD pruning, data post-filtering strategy, algorithm, etc.) on the final results variability (Fig. 2).

Results

Table 1: % results overlap between significant SNP pairs under index of 3 and non-filtered data (exhaustive)

<table>
<thead>
<tr>
<th>MB-MDR</th>
<th>CODOMINANT</th>
<th>ADDITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>with LD</td>
<td>without LD</td>
<td>with LD</td>
</tr>
<tr>
<td>97.5</td>
<td>82.1</td>
<td>40.3</td>
</tr>
<tr>
<td>59.3</td>
<td>88.9</td>
<td>44.4</td>
</tr>
<tr>
<td>25.9</td>
<td>31</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Note: The higher % result overlap, the lesser effect given variable has on results consistency

Table 2: % results overlap between MB-MDR methods run on pre-filtered marker set using AS related set or implication index of 3

<table>
<thead>
<tr>
<th>MB-MDR</th>
<th>CODOMINANT</th>
<th>ADDITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>with LD</td>
<td>without LD</td>
<td>with LD</td>
</tr>
<tr>
<td>22.2</td>
<td>27.3</td>
<td>64.2</td>
</tr>
<tr>
<td>16.7</td>
<td>27.3</td>
<td>43.2</td>
</tr>
<tr>
<td>42.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The higher % result overlap, the lesser effect given variable has on results consistency

Conclusions

- It was again confirmed that LD effect can lead to “redundant epistasis” and/or negatively affect the final results consistency. For example, compare BOOST exhaustive against MB-MDR ADDITIVE comparison (4 cells) (Table 1)
- BOOST is best compatible with MB-MDR co-dominant run on LD pruned data (Table 1) also confirmed by rank analysis (Fig. 2)
- The pre-filtering based on implication index is a better strategy compared to more restrictive gene lists resulting in 2x fold increase in % overlap due to larger set of candidate SNP pairs obtained via imp. index approach (Table 2)
- MB-MDR run in co-dominant mode on LD pruned data (without LD) provides the highest robustness with respect to LD pre-filtering and main effects correction model (Table 2)
- Our preliminary results from MB-MDR non-exhaustive analysis show that co-dominant lower-order effects correction scheme in MB-MDR seems to be less susceptible to LD effects compared to the additive one (Fig 2).

References


Figure 1: GWAS workflows used in this study testing the effects of various variables (LD status, marker pre-filtering strategies, low-order correction scheme) with application of BOOST and MB-MDR algorithms. [*] an optimized version of BOOST like implementation identical to original algorithm but accounting for missing genotype values was used

Figure 2: Dendrogram and its distance matrix comparing variability of selected workflows using ranks of common 36 SNP pairs. The results are based on pre-filtered data under implication index of 3 (Methods)