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 pathology and approaches us closer to the ultimate goal - personalized medicine.

In this work we rely on a minimal GWAI 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 BOolean Operation-based Screening and Testing (BOOST) algorithms [4][*] for detection of statistically significant epistatic SNP-SNP interactions, we investigate the effect of exhaustive (BOOST) and non-exhaustive (MB-MDR) marker processing strategies, 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 - 13 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), algorithm 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 Golden Helix with LD correlation threshold of 0.75 and window size of 50 bp with 1 bp increment.

Data filtering using Biofilter 2.0

The search-space was reduced to maximize chances of finding truly biologically relevant SNPs (i.e. pre-filtering). This was done using Biofilter 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*, *ERAP1* and *KIF21B* [2].

% overlap between workflows

The final results represented as a list of significant SNP pairs with corresponding statistics were compared across workflows (**Fig.1**). The maximal %overlap value between final results (*maximum*[(# of common SNPs pairs/# of total SNP pairs **workflow 1**) , (# of common SNPs pairs/# of total SNP pairs **workflow 2**)]) from 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.

Eucledian distance between workflows was found using an input vector of 36 ranks of common SNP pairs from top 1000 in the final MB-MDR results list. The resulting dendrogram (hierarchical tree) shows the impact of different choices (LD pruning, data pre-filtering strategy, algorithm, etc.) on the final results variability (**Fig. 2**).

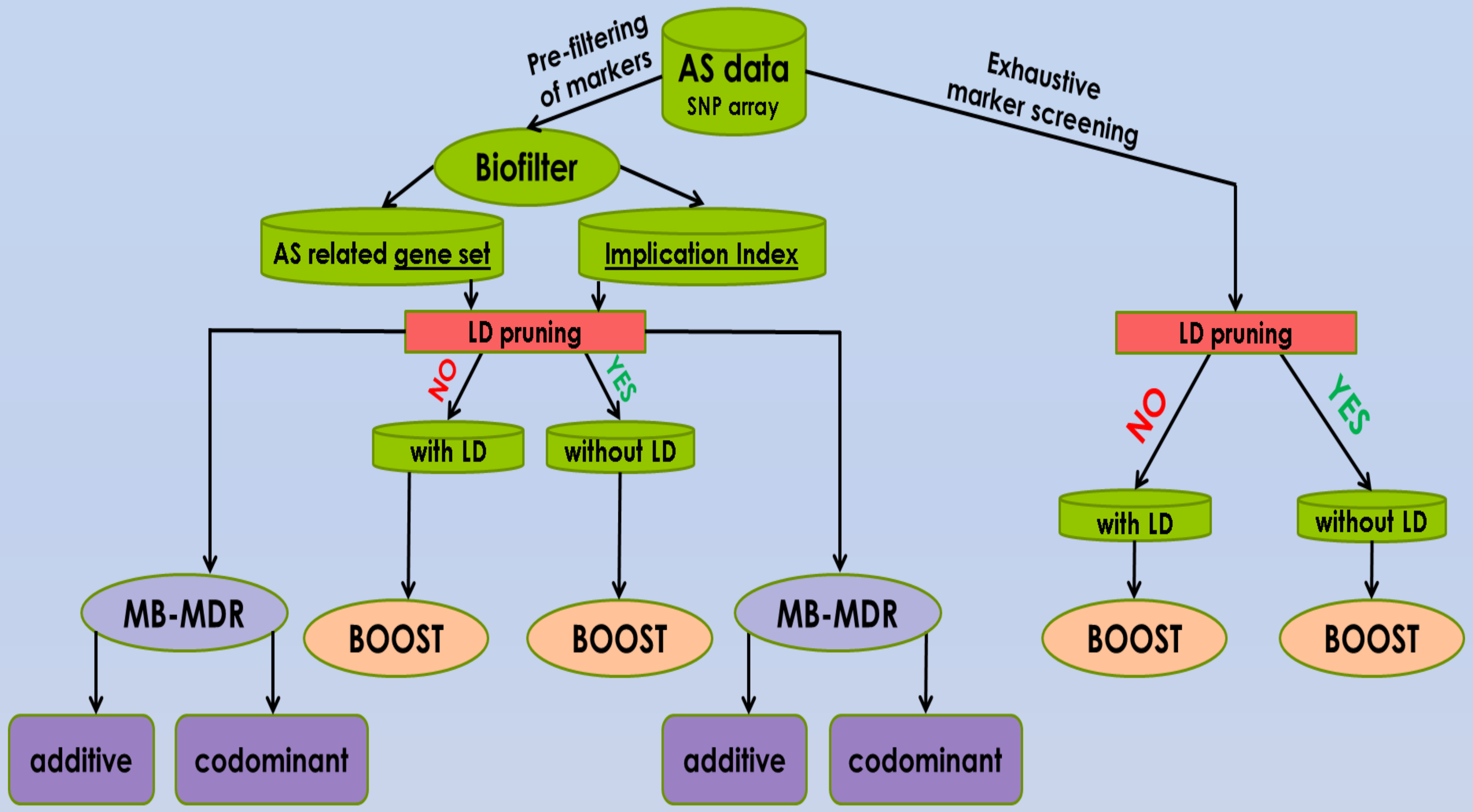
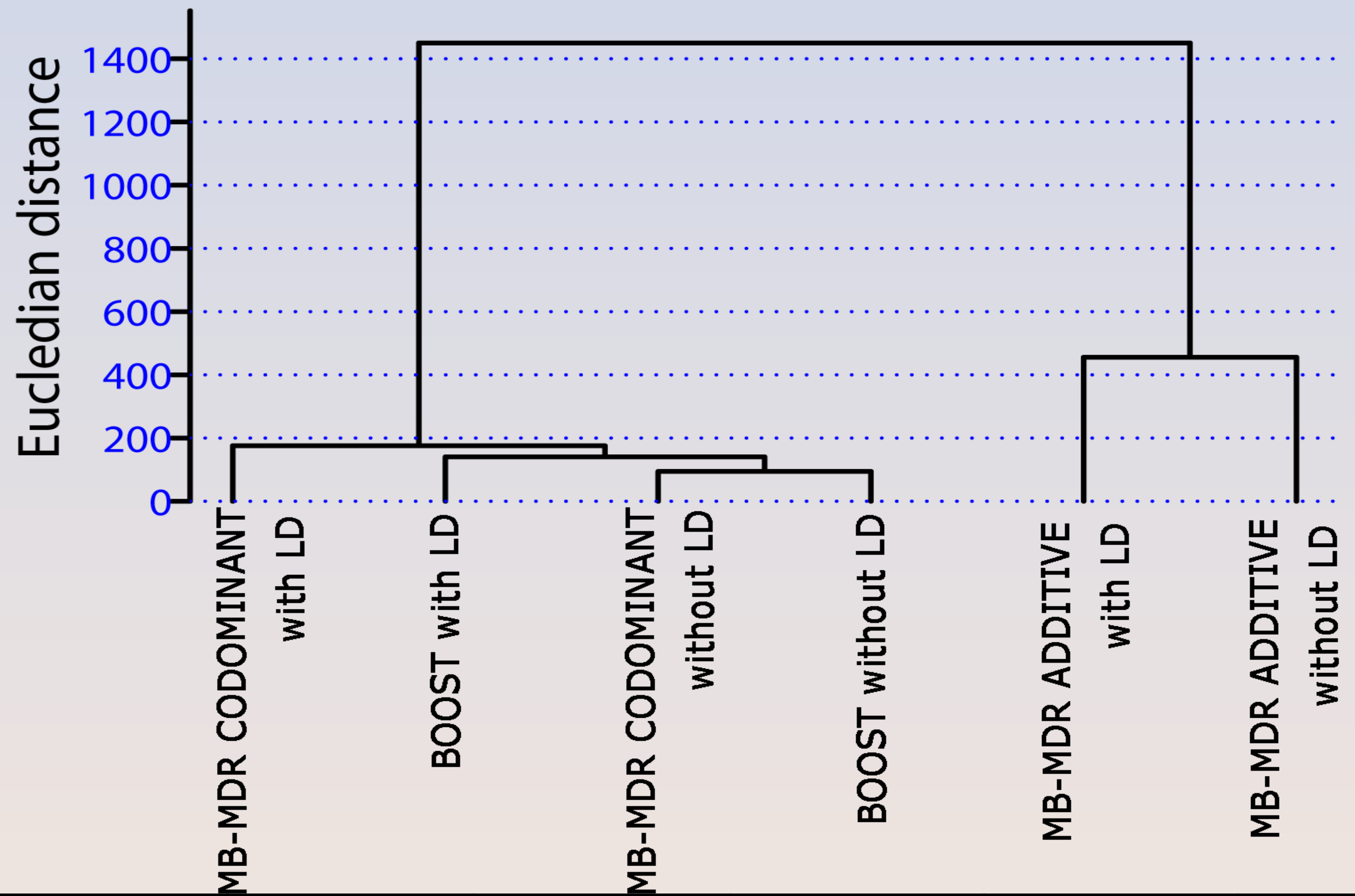


Figure 1: GWAI workflows used in this study testing the effects of various variables (LD status, marker pre-filtering strategies, low-order correction scheme) with application of BOOST and MB-MDR algorithms. [*] an optimized version of BOOST like implementation identical to original algorithm but accounting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



	MB-MDR ADDITIVE with LD	MB-MDR CODOMINANT with LD	MB-MDR ADDITIVE without LD	MB-MDR CODOMINANT without LD	BOOST with LD
MB-MDR CODOMINANT with LD	1401.5				
MB-MDR ADDITIVE without LD	**455.8**	957.9			
MB-MDR CODOMINANT without LD	1442.9	106.3	992.2		
BOOST with LD	1406.7	156.0	961.2	140.7	
BOOST without LD	1450.1	175.6	998.6	**94.4**	100.1

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**)

Results

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

			MB-MDR			
			CODOMINANT		ADDITIVE	
			pre-filtered			
			with LD	without LD	with LD	without LD
BOOST	pre-filtered	with LD	97.5	82.1	40.3	27.6
		without LD	59.3	**88.9**	44.4	**45.7**
	exhaustive	with LD	13.6	4.5	13.6	9.1
		without LD	25.9	**31**	1.4	**1.4**

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 gene set or Implication Index of 3

				MB-MDR			MB-MDR	
				CODOMINANT			CODOMINANT	
				with LD	without LD		with LD	without LD
				Gene set			Imp. Index	
MB-MDR	ADDITIVE	with LD	Gene set	22.2	27.3	Imp. Index	64.2	41.7
		without LD		16.7	**27.3**		43.2	**42.9**

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 hightest 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

[1] Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K: "**Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise.**" *Ann Hum Genet* 2011, 75:78-89.

[2] Evans DM, et. al. "**Integration between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility**." *Nature Genetics* 2011, Sep;43(9):919

[3] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, "**Potential etiologic and functional implications of genome-wide association loci for human diseases and traits"**. *Proc Natl Acad Sci USA*. [May 27, 2009]

[4] Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. "**BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies"**. Am J Hum Genet. 2010 Sep 10;87(3):325-40

[5] Braem K, Lories RJ. "**Insights into the pathophysiology of ankylosing spondylitis: contributions from animal models"**. Joint Bone Spine. 2012 May;79(3):243-8

[6] Gusareva ES, Huyghe JS, Van Steen K. **Genome-wide epistasis screening for asthma associated traits. Joint statistical Meetings 2011 conference**. Abstract number #301420. Miami Beach, Florida.

[7] Mahachie John, Jestinah et. al. **A robustness study to investigate the performance of parametric and non-parametric tests used in Model-Based Multifactor Dimensionality Reduction Epistasis Detection**. *The Capita Selecta in Complex Disease Analysis (CSCDA) 2012 second edition conference*.

[8] Mahachie John JM, Cattaert T, Lishout FV, Gusareva ES, Steen KV. **Lower-order effects adjustment in quantitative traits model-based multifactor dimensionality reduction**. PLoS One. 2012;7(1)

[9] Gusareva et al. 2009. **GENOME-WIDE ASSOCIATION INTERACTION ANALYSIS FOR COMPLEX DISEASES: an example on Alzheimer disease**. Submitted Feb 2013

[10] William S. Bush, Scott M. Dudek, Marylyn D. Ritchie. **Biofilter: A Knowledge-Integration System for the Multi-Locus Analysis of Genome-Wide Association Studies**. Pac Symp Biocomput. 2009: 368–379