A cautionary note on the impact of protocol changes for Genome-Wide Association SNP x SNP Interaction studies: an example on ankylosing spondylitis

Kyrylo Bessonov (✉) · Elena S. Gusareva · Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, University of Liège, Quartier Polytech 1, Allée de la Découverte 12, 4000 Liège, Belgium

Systems Biology and Chemical Biology, GIGA-R, University of Liège, Avenue de l'Hôpital 11, 4000 Liège, Belgium

e-mail: kbessonov@ulg.ac.be

Abstract Genome-wide association interaction (GWAI) studies have increased in popularity. Yet to date, no standard protocol exists. In practice, any GWAI workflow involves making choices about quality control strategy, SNP filtering, linkage disequilibrium (LD) pruning, analytic tool to model or to test for genetic interactions. Each of these can have an impact on the final epistasis findings and may affect their reproducibility in follow-up analyses. Choosing an analytic tool is not straightforward, as different such tools exist and current understanding about their performance is based on often very particular simulation settings. In the present study, we wish to create awareness for the impact of (minor) changes in a GWAI analysis protocol can have on final epistasis findings. In particular, we investigate the influence of marker selection and marker prioritization strategies, LD pruning and the choice of epistasis detection analytics on study results, giving rise to 8 GWAI protocols. Discussions are made in the context of the ankylosing spondylitis (AS) data obtained via the Wellcome Trust Case Control Consortium (WTCCC2). As expected, the largest impact on AS epistasis findings is caused by the choice of marker selection criterion, followed by marker coding and LD pruning. In MB-MDR, co-dominant coding of main effects is more robust to the effects of LD pruning than additive coding. We were able to reproduce previously reported epistasis involvement of *HLA-B* and *ERAP1* in AS pathology. In addition, our results suggest involvement of *MAGI3* and *PARK2,* responsible for cell adhesion and cellular trafficking. Gene Ontology (GO) biological function enrichment analysis across the 8 considered GWAI protocols also suggested that AS could be associated to the Central Nervous System (CNS) malfunctions, specifically, in nerve impulse propagation and in neurotransmitters metabolic processes.

Keywords Genome-wide association interaction (GWAI), epistasis, protocol adoption, ankylosing spondylitis

Introduction

High-throughput technologies give access to unprecedentedly vast amounts of data such as Single Nucleotide Polymorphisms (SNPs). In Genome Wide Association Studies (GWAS), thousands of these are scanned for their potential association with traits of interest, such as a disease status. Hard to disentangle are complex traits which assume an intricate interplay between genetic, environmental and/or many other unknown factors. For these traits added benefits can be obtained by using methods that account for biological and statistical interactions, rather than by adopting strategies that analyze each SNP at a time. This is the subject of Genome-wide association interaction (GWAI) studies, which usually focus on pairwise SNP x SNP interactions. It is believed that GWAI studies can lead to novel or improved clinical and biologically relevant hypotheses.

Many strategies exist to carry out a GWAI study, such as those based on generalized linear regression models (GLM), BOOST (Wan et al. 2010), Dimensionality Reduction (MB-MDR) (Cattaert et al. 2011; Van Lishout et al. 2013), MDR (Ritchie et al. 2001), BiForce (Gyenesei et al. 2012), Bayesian Models (e.g. BEAM) (Zhang et al. 2011) and several others (Pang et al. 2013; Van Steen 2012; Wei et al. 2014b; Zhang et al. 2008). For extensive reviews, please refer to (Gusareva and Van Steen 2014; Van Steen 2012; Wei et al. 2014a). All of these methods have their pros and cons, but the problems or hurdles encountered during the analysis are largely overlapping. Common hurdles to overcome include dealing with high dimensionality, handling a huge multiple testing problem, limiting computation time (when assessing statistical significance), and controlling false positive rates (Van Steen 2012). Unfortunately, often when novel GWAI analysis methods are introduced the impact on epistasis findings of changes in the GWAI protocol are given limited attention. Some examples of key protocol parameter changes relate to marker filtering/prioritization, LD thresholds in marker pruning, *a priori* assumptions about operating two-locus inheritance models, main effects correction. It is essential to differentiate between global two-locus testing (i.e. not differentiating between main effects and interaction effects) and specific interaction testing (i.e., testing for the interaction between two loci itself, above and beyond the main effects). Specific interaction testing requires making adjustments for lower-order effects, and hence proposing a particular encoding scheme for lower-order effects. Several authors have commented upon the limitations of an additive encoding scheme for SNPs in SNP x SNP interaction studies and recommended co-dominant coding (Mahachie John et al. 2011b)

In this study, we investigated the impact on final epistasis results of changing one or more parameter settings in a GWAI protocol, leading to 8 interesting strategies (Fig. 1 and Table S1). These strategies are motivated by prior theoretical work (Cattaert et al. 2011; Grange 2014; Mahachie John et al. 2012). As a benchmark protocol, we took the one proposed by (Gusareva and Van Steen 2014). As analytic tools we chose BOOST (Wan et al. 2010), motivated by its popularity and computational efficiency due to a Boolean data representation, and MB-MDR (e.g., Cattaert et al. 2011), because of its non-parametric nature regarding epistasis models and its ability to correct for confounders or lower-order effects. In brief, BOOST handles binary traits and fits a full generalized linear model with main SNP effects (2 degrees of freedom (df) for each main effect) and SNP x SNP interaction effects (4 df). Significant (specific) interactions are identified via a Log-Likelihood Ratio Test (LRT) based on 4 df. The

Bonferroni correction is proposed as a multiple testing corrective measure. In contrast, MB-MDR handles binary, continuous, and censored traits, and first carries out a dimensionality reduction procedure while pooling risk-alike multi-locus genotype combinations together. Its final test statistic contrasts high risk versus low risk multi-locus genotypes. While correcting for multiple testing, significance is assessed via the resampling based strategy proposed by (Westfall 1993). For additional details about MB-MDR and BOOST, we refer to (Cattaert et al. 2011; Mahachie John et al. 2012; Van Lishout et al. 2013; Wan et al. 2010). To achieve our goal, we used real-life ankylosing spondylitis (AS) data from the Wellcome Trust Case Control Consortium (WTCCC2). Ankylosing spondylitis (AS) is a common form of inflammatory arthritis occurring in approximately 1 to 14 out of 1,000 adults globally (Dean et al. 2014). Apart from confirming previously known AS associated genes (Alvarez-Navarro and Lopez de Castro 2013; Evans et al. 2011), we will show that combining different protocols may give new insights into disease pathology.

Materials and Methods

Data Quality Control

Approved access to Wellcome Trust Case Control Consortium (WTCCC2) data, in particular via EBI accession no. EGAS00000000104, EGAD00010000150, EGAD00000000024 and EGAD00000000022, resulted in a dataset composed of 2005 Ankylosing Spondylitis (AS) cohort samples, and 3000 British 1958 Birth Cohort (BC) and 3000 National Blood Donors (NBS) Cohort samples. The 1788 cases were of British Caucasian origin recruited by Nuffield Orthopedic Centre, Oxford and Royal National Hospital for Rheumatic Diseases, Bath. The first batch of case samples were genotyped on an Illumina 670k platform, the last two batches of control samples were genotyped on an Illumina 1.2M platform. No imputation was done for these genotypes. We used PLINK (Purcell et al. 2007) to select 6,587 subjects (1788 cases plus 4799 controls), 3409 of which were male and 2864 female, and 487,780 SNPs, according to criteria described in (Evans et al. 2011). Briefly, SNPs showing MAF < 0.01, Hardy-Weinberg *p*-values $< 5 \times 10^{-20}$ and SNPTEST information measure < 0.975 were excluded. The dataset inflation factor (λ) was estimated as 1.02917. The QC-ed genotype data were stored in GEN format and were converted to PED and MAP files using GTOOL from Oxford University, UK (Colin Freeman 2012).

Additional data handling

Depending on the GWAI protocol of choice, additional data manipulations were required, such as marker prioritization or LD pruning (Fig. 1). We prioritized markers with the Biofilter 2.0 software developed by Ritchie et al. (Bush et al. 2009). The Biofilter 2.0 uses a list of public biological databases (sources) such as KEGG, BioGRID, MINT, via the Library of Knowledge Integration (LOKI), to generate pairwise gene-gene interaction models (Wan et al. 2010). No disease specific information was used, but available knowledge about gene-gene interactions from different biological resources called by Biofilter 2.0 (Bush et al. 2009). The advantage of such an approach is an 11-fold reduction of the original marker set, without selection bias introduction towards a particular disease. The disadvantage of *any* pre-filtering method is that useful information may be disregarded and biologically relevant

SNPs removed from further analysis protocols. In practice, taking the 487,780 SNPs from (Evans et al. 2011) as a starting point, we applied Biofilter 2.0 with a minimum implication index threshold of 3, meaning that at least 3 data sources confirmed the associated gene-gene interaction. This resulted in the generation of 8,288 gene-gene models and a set of 44,018 unique SNPs (Fig. 1).

To reduce the number of tests and the number of false positives based on genomic proximity (for instance, redundant epistatic SNP pairs), some GWAI protocols involve LD filtering or pruning (Fig. 1). As motivated and recommended by (Gusareva and Van Steen 2014), we adopted a rather mild pruning threshold of r^2>0.75, still allowing for moderate LD but removing strong LD. Pruning at r^2>0.75 threshold implies that every SNP pair in the pruned dataset has an r^2 of at most 0.75. The proposed threshold offers a balance between power gain and false positives due to high LD. In practice, LD-pruning was performed considering the sliding windows of size 50 (i.e., 50 markers) with window increments of 1 marker. For any pair of markers under testing whose r^2>0.75, the first marker of the pair was discarded, as implemented in SVS Version 7.5 (Golden Helix, Inc.) (Bozeman 2015). After LD pruning, the original marker dataset reduced from 487,780 to 321,565 markers. After LD pruning, the biofiltered data (Biofilter 2.0) reduced from 44,018 to 30,426 markers (Fig. 1).

Interaction testing

To test for interactions we used two software tools: BOOST (Wan et al. 2010) and MB-MDR (Cattaert et al. 2011). We extended the original BOOST algorithm as it did not deal with missing genotypes and so as to properly adjust the number of degrees of freedom (df) in case less than 3 genotypes was observed for a marker. Our implementation of BOOST was coded in C++ and can be obtained upon request, via the corresponding author. Notably, a similar adaption of BOOST was implemented in the PLINK software (PLINK version 1.9, called via "--fast-epistasis boost"). In practice, for the MB-MDR methodology, we used the algorithms implemented in MBMDR version 3.0.2 (Van Lishout et al. 2013) that provides several advantages over classic MDR (Ritchie et al. 2001) or BOOST, such as the ability to analyze different trait types within the same framework, as well as non-parametric model free testing for two or three-order interactions while adjusting for lower order effects or relevant confounders. Since MBMDR versions 2.0 – 4.1.0 require significant computational resources to run on a genome-wide scale, we were not able to use these MB-MDR versions on unfiltered data, at the time of analysis. The version that allows for exhaustive genome-wide epistasis screening is underway. Hence, in this study, all MB-MDR based protocols (Fig. 1) were implemented on a reduced dataset via Biofilter 2.0. The default main effects correction in MB-MDR is a co-dominant one. As was mentioned in (Mahachie John et al. 2011b), it is important to correct for main effects in a co-dominant way to avoid false epistasis signals.

Results obtained from either one of the 8 GWAI protocols included in this study were compared to results obtained in the reference study (Evans et al. 2011). In particular, as statistical interactions may be indicative for important main effects (Greene et al. 2009), we compared SNPs derived from significant SNP pairs to the list of 49 SNPs in Supplementary Table S2 of (Evans et al. 2011) that passed quality control in their replication analysis. Also, significant SNP pairs obtained in this work were compared to the reference panel of 102 SNP x SNP pairs tabulated in Supplementary Table 5 of (Evans et al. 2011). The latter table lists all considered SNP pairs for interaction testing,

using an additive x additive term in a logistic regression model (i.e. additive encoding of SNP main effects and interaction).

Assessing consistencies between protocols

The overlap between GWAI protocols (Fig. 1) in identifying the same significant SNP pairs was graphically presented via the Euler diagram (Fig. 2) with the software VennMaster 0.38 (Kestler et al. 2005). For each of the SNP pairs tested, ranks were computed, for each protocol separately, with rank 1 assigned to the SNP pair with the smallest multiple testing corrected *p*-value. Then, SNP pairs that were common to each protocol were retained, in order to be able to compare exhaustive with non-exhaustive protocols. A total of 1230 SNP pairs were retained. These are listed in Table S4, together with their associated protocol-specific *p*-values, and were subsequently used to calculate "distances" between protocols. In particular, we calculated the squared Euclidean distance between 8 GWAI protocols using 8 input vectors containing 1230 ranks each. These 1230 ranks for each protocol corresponded to relative positions of the common 1230 SNP pairs amongst all ordered SNP pairs (from highest to lowest significance). For example, the ranks for the *rs12026423* x *rs7528311* pair in protocols 1 to 8 were 232, 2300, 97, 61, 259, 151, 59892 and 43598, respectively. We used *complete linkage* cluster agglomeration with *hclust()* to build a dendrogram (hierarchical tree) (RCoreTeam 2013) (Fig. 3). The use of SNP pair ranks coupled with hierarchical clustering allows an unbiased qualitative comparison of the top findings derived via different GWAI protocols.

In addition, to assess the effects of MAFs on top findings in each protocol, we selected the top 1000 SNP pairs for each GWAI protocol. We subsequently defined the following MAF classes or bins, using interval notations: 1) (0-0.05) (MAF<0.05; less common minor allele); 2) [0.05-0.10) ($0.05 \leq$ MAF < 0.10; moderate occurrence of the minor allele); 3) [0.10-0.50) ($0.10 \leq$ MAF < 0.50; rather common minor allele). Two-dimensional bins were defined by combining the aforementioned three 1-dimensional bins as follows: 1) (0-0.05)/(0-0.05); 2) [0.05-0.10)/(0-0.05); 3) [0.10-0.50)/(0-0.05); 4) [0.05-0.10)/[0.05-0.10); 5)[0.05-0.10)/[0.05-0.10); 6)[0.10-0.50)/[0.10-0.50). Note that for any SNP pair falling into one of these six 2-dimensional bins, the MAF of the first SNP in the pair will be larger or equal than the MAF of the second SNP in the pair, unless perhaps when both SNPs belong to the same one-dimensional bin.

Biological relevance

The SNP to gene symbol annotation (when possible) was done using SCAN – a SNP and CNV Annotation Database (Gamazon et al. 2010) The SCAN database accepts a list of SNPs, maps them to genomic coordinates and outputs corresponding gene symbols, provided that the SNP is located within a gene coding region, which is helpful in assessing putative biological function and context. We then performed GO enrichment analyses (Huang da et al. 2009) on the top 1000 most significant SNP pairs, by GWAI protocol. In practice, we used the *topGO* library in R that takes into account the GO graph structure and removed nodes (GO terms) that had a low number of annotated

genes, i.e., less than 10 (Ackermann and Strimmer 2009; Alexa et al. 2006). The *weight01* algorithm was chosen based on the author's recommendations and due to shared benefits of the *elim* and *weight* algorithms (Ackermann and Strimmer 2009). Significance of each GO term, per protocol, was based on Fisher's exact test. Overall significance across all protocols was assessed via Fisher's combined probability test at a significance level of 0.05

Results

Consistency between interaction results derived from different GWAI protocols

A graphical representation, showing the overlap of significant findings between considered GWAI protocols is presented in Fig. 2. The significant SNP pairs (multiple testing corrected) retrieved via GWAI protocol #1-#8 (Fig. 1) are tabulated in Table S3. The largest number of significant SNP pairs were obtained for protocols that use additive encoded corrections for main effects (protocols #7, #8). Over 2000 significant pairs were detected with an exhaustive implementation of BOOST on LD-pruned data (protocol #2). The number of significant SNP pairs reduces significantly when BOOST is used exhaustively on un-pruned data (protocol #1; 226 pairs). All other protocols identified less than 130 siginficant epistasis signals; the most liberal is protocol #3 (BOOST on filtered data), the most conservative is protocol #6 (MB-MDR on biofiltered and LD-pruned data), also using a co-dominant encoding scheme to correct the interaction testing for lower order SNP effects. Furthermore, only few of the findings obtained via exhaustive protocols (BOOST, #1-#2) were retrieved via protocols that first biofiltered the data (protocols #3-#8). With the same protocol for LD pruning on biofiltered data, both BOOST and MB-MDR in co-dominant main effects correction mode, gave partially overlapping results (Fig. 2). In effect, over 97% of significant SNP x SNP interactions identified via MB-MDR protocols #5 and #6 were identified in BOOST protocols #3 and #4, respectively (Fig. 2 and Table S3).

Via hierarchical clustering (see Methods for details), the largest distance between protocols (i.e., the smallest overlap between top findings, not necessarily significant) was obtained for exhaustive screening protocols: protocol #1 - BOOST without pruning and protocol #2 – BOOST applied on LD-pruned data (Fig. 3). The effect of LD in BOOST applications is less pronounced when data were first biofiltered. Actually, the smallest distance between protocols was observed between protocols #3 (BOOST without LD pruning) and #4 (BOOST applied to LD-pruned data). In general, the effect of LD on SNP pair rankings seems to be smaller in non-exhaustive protocols as compared to the exhaustive protocols considered. The second smallest distances observed between protocols was between #5 and #6 (MB-MDR with co-dominant correction of lower-order effects) and between #7 and #8 (MB-MDR with additive encoding of main SNP effects). Within non-exhaustive screening protocols (#3-#8), analyses that used an additive encoding to adjust for SNP main effects while testing for interactions stood out; all protocols involving epistasis detection analytics with co-dominant encoding schemes of some sort clustered together (Fig. 3). A closer look at the overlapping significant SNP pairs across all 8 GWAI protocols, reveals that only 3 out of 1230 SNP pairs

(*rs12026423/rs7528311*, *rs11964796/rs13194019* and *rs13194019/rs1784607*) met statistical significance at α=0.05, according to at least one GWAI protocol (Table 1 and S4).

We furthermore investigated whether any of the 49 main effects SNPs reported in (Evans et al. 2011) were supported by our SNP x SNP interaction results across the 8 tested GWAI protocols (see Methods for more details). With GWAI protocols #5, #6, #7 and #8 based on the MB-MDR framework, we were able to confirm *rs9788973* (*p*-value 0.49), which maps to *HLA-B* and *rs30187* (*p*-value 1.1×10^{-9}), which maps to *ERAP1* (Evans et al. 2011). These SNPs occurred in the pairs *rs2523608 x rs9788973* and *rs30187 x rs284498* (see Table 2). Only GWAI protocols #7 and #8 coined the aforementioned two pairs as being statistically significant. None of the 102 SNP pairs listed in (Evans et al. 2011) were found to be statistically significant in our re-analysis, regardless of the protocol used. Relaxing the conditions, we determined the number of SNP pairs with a SNP that occurred in at least one of the 102 SNP pairs reported by (Evans et al. 2011). A total of 38 such SNP pairs could be detected. These are listed in Table S5. From these, only 8 significant SNP pairs were highlighted by at least one of our GWAI protocols (in particular, protocol #7 and #8 - Table 3)

To investigate the influence of MAFs on epistasis findings using different protocols, we defined six 2-dimensional bins (see Methods for more information). The allocation of top 1000 epistasis findings (significant or not) to either of these bins is presented in Fig. 4. Hence, adding up the number of allocated SNP pairs to each bin (red numbers in Fig. 4), within the same protocol, gives 1000. Within the exhaustive protocols (#1 and #2, respectively BOOST applied to unpruned and LD-pruned data), there is a tendency for SNP pairs each having MAF $\geq$ 0.05 to occur in the top 1000. The same is observed for non-exhaustive protocols that rely on additive encodings when adjusting for main effects (protocols #7 and #8, MB-MDR applied to unpruned and LD-pruned data, respectively). The highest number of SNP pairs (out of 1000) with MAFs < 0.05 were obtained with exhaustive BOOST screening on unfiltered and unpruned data (protocol #1). In general, all protocols give rather similar results, apart from protocols with additive main effects correction (#7 and #8, MB-MDR) for which virtually all of the top 1000 SNP pairs involved at least one SNP with MAF$\geq$0.10 (respectively, 100% and 100%). For protocols #1-#6, the percentage of SNP pairs appearing in the top 1000 list with at least one MAF < 0.05 ranged from 0.2% (protocol #2) to 5.9% (protocol #1).

Biological relevance

To provide a biological context, we performed a GO functional enrichment analysis on the top 1000 SNP pairs identified within each individual GWAI protocol. Each SNP was mapped to a gene, when possible (see Methods for additional details). A GO term was considered when at least 10 of these genes could be annotated to them. This led to a total of 480 common GO terms across all 8 GWAI protocols with combined *p*-values < 0.05 (Table S6). Top 10 GO terms are shown in Table 4. Using a significance level of 0.05, significant combined *p*-values were obtained for GO terms related to the central nervous system (CNS). In particular, links between AS pathology and nervous system signal transmission via synapses biological processes was observed via e.g. GO:0007411 (combined *p*-value: 7.86×10^{-77}), GO:0007268 (combined *p*-value: 2.00×10^{-36}), and GO:0043524 (combined *p*-value: 2.91×10^{-17}). To a lesser degree, we also observed a link between AS and immune system processes that involve antigen processing and

presentation via MHC complex: combined *p*-value for GO:0002479 of 1.77×10^{-8} (not corrected for multiple testing). Other overall significant GO terms were linked to biological processes such as membrane transport (GO:0055085, combined *p*-value: 3.04×10^{-50}) and sudden response to stimuli (GO:0001964, combined *p*-value: 1.48×10^{-10}) without a clear association to AS. In addition, we detected an involvement of the *Notch* pathway responsible for the proliferation of neurons (GO:0007219, combined *p*-value of 1.02×10^{-5}), again linking AS to CNS processes.

Discussion

In our study, we demonstrated that choices about data filtering, pruning and lower order effects adjustment may cause substantial variation in epistasis findings. We demonstrated this by making changes to the reference GWAI protocol we published earlier (Gusareva and Van Steen 2014), giving rise to 8 GWAI protocols under investigation in this work (Fig.1). The reference GWAI protocol consists of a set of guidelines designed to address problems of epistasis reproducibility in the context of genome-wide epistasis screening with thousands of SNP markers. It contains recommendations on rigorous data quality control steps, exhaustive or non-exhaustive marker screening, LD pruning thresholds and the selection of a suitable analytic epistasis detection tool.

Based on our results (for instance Fig 2) the major cause of heterogeneity in findings is the choice about which markers to retain in the analysis. We referred to it as "pre-selection of markers". We used filtering based on biological knowledge to make educated pre-selections, using a compendium of biological databases via Biofilter 2.0 (Bush et al. 2009). The effects of pre-selections on the number of SNPs can be huge, as was exemplified on AS: before selection, 487,780 SNPs; after selection, 44,018 SNPs. This has huge consequences for subsequent analyses. In a negative sense, there is a risk of removing pairs of SNPs that may lead to interesting new hypotheses, for which no reported evidence exists in existing biological data repositories. In a positive sense, less multiple tests are need to be performed, hereby reducing computation time and potentially also the number of false positives. Seeking a balance between potentially improving the power of the GWAIs by relying on prior knowledge versus decreasing the chance of missing important findings remains a challenging task. When inspecting the overlap between significant results for each protocol, it is therefore not surprising that little overlap may exist between significant results obtained via exhaustive protocols and significant results obtained via non-exhaustive protocols. In fact, for the AS data we re-analyzed, no overlap was found at the SNP level (see Fig. 2 and Fig. 3 protocols #1-#2 versus #3-#8). Furthermore, the protocol adopted by (Evans et al. 2011) makes a heavy pre-selection of markers. Only those SNPs showing a significant association with AS via main effects GWAs were considered. This involved 15 SNPs, half of which were also included in the 487,780 SNPs that served as input to our own GWAI protocols (#1-#8): *rs30187, rs10781500, rs10865331, rs11209026, rs2297909, rs378108, rs11209032*. The likelihood ratio interaction tests adopted in their work were similar to the ones implemented in BOOST. However, whereas in BOOST tests are based on 4df, interaction tests in (Evans et al. 2011) were based on 1df (testing departure from additivity on the log-odds scale). Hence, it is not surprising that none of the significant SNP pairs reported in (Evans et al. 2011) could be reproduced in our study. Notably, neither BOOST nor MB-MDR in our protocols adjusted for population stratification. In contrast, (Evans et al. 2011) did correct for potential population stratification using a two-stage approach involving Bayesian clustering and Hidden Markov models. In theory, this may explain additional

differences between our analyses and the ones performed in the reference study (Evans et al. 2011). However, given that the inflation factor based on median X^2 for the AS data is 1.02917, we believe that no adjustments were necessary and hence no spurious results were generated as a result of not correcting for population stratification in our adopted protocols.

Our results, visualized in Fig. 3, suggest that the second largest cause for heterogeneity in significant findings, derived from different protocols, is the adopted encoding scheme for genetic variants. This is clear for the non-exhaustive protocols included in our study (#5-#8). It is less clear for exhaustive protocols, since the ones included in our study only considered co-dominant encoding schemes (#1-#2). However, our experience with other real-life applications seems to support our suggestion also for exhaustive protocols (data not shown). Previous theoretical work also showed that additive encodings for lower order effects may increase false positives rates in interaction studies (Mahachie John et al. 2012). This is in line with the large number of significant interactions identified via protocols #7 and #8 (Fig. 2). It is very unlikely that over 50000 significant interactions highlighted by these protocols are genuine, and are caused by the (strong) main effects blurring the epistasis signal (Mahachie John et al. 2012).

The third largest cause for heterogeneity is attributed to differences in employed LD-pruning approaches. Here, the effect of LD-pruning (i.e., pruning at $r^2 < 0.75$ or not) was more pronounced under additive encoding schemes (protocols #7 versus #8) as opposed to co-dominant encoding strategies (protocols #3 versus #4, and protocols #5 versus #6). Therefore, it is important to discuss the primary interaction study performed in (Evans et al. 2011), targeting additive x additive interactions, with caution, and in the light of the adopted pruning protocol. Fig. 3 shows that the effects of LD pruning are more severe for exhaustive protocols compared to non-exhaustive protocols. This is not surprising, given that the LD pruning in the first implies a reduction of about 150,000 SNPs, compared to less than 15,000 SNPs in the second. Hence, although potentially more significant SNP pairs can be revealed in protocol #1 (exhaustive, BOOST, unpruned), less significant pairs are highlighted as compared to protocol #2 (exhaustive, BOOST, LD-pruned; Fig. 1). This can be explained by the reduced number of tests to account for Bonferroni corrections. The reverse is observed for protocols #3 (BOOST, pre-selected) and #4 (BOOST, pre-selected and LD-pruned). Here, protocol #4 gives rise to less significant SNP pairs compared to protocol #3 (Fig.2). There is still a reduction of the multiple testing burden in protocol #4 is true, but this cannot explain the phenomenon. More likely, an increased number of redundant epistasis signals (due to high LD between some marker pairs) are an explanatory factor. The same can be observed for MB-MDR-based protocols #5 and #6. In particular, again LD pruning as part of protocol #6 gives rise to a smaller number of significant SNP x SNP interactions (47 – see Fig. 2) compared to protocol #5 (no LD pruning; 77 – Fig. 2). Note that MB-MDR and BOOST use quite different multiple testing correction strategies. In case of BOOST, a conservative Bonferroni correction is advocated. In MB-MDR, a permutation-based *maxT* strategy is implemented, which relies on subset pivotality to guarantee strong FWER control at $\alpha = 0.05$.

Less common and rare variants tend to increase false positive rates, when inappropriate tests are used, as reported in (Mahachie John et al. 2011a; Tabangin et al. 2009). According to (Tabangin et al. 2009) rare SNPs with MAF < 0.05 showed a significantly higher likelihood of being classified as false positives in the logistic regression based GWAS (Tabangin et al. 2009). For BOOST-based protocols (#1 - #4), the percentage of top 1000 SNP pairs with at least one MAF < 0.05 that were statistically significant (multiple testing corrected), was respectively 5.9%, 0.2%, 4.9 % and

2.4% (data not shown). For MB-MDR based protocols (protocols #5-#6) the percentage of such SNP pairs was respectively 0.1% and 0.2%, smaller than with BOOST-based protocols. However, for MB-MDR based protocols #7 and #8 (using additive encoding schemes for main effects adjustment), the percentages were higher (4.8% and 5.3%, respectively). This is in line with earlier findings about MB-MDR performance (Mahachie 2012; Mahachie John et al. 2012; Mahachie John et al. 2011b). When MB-MDR is applied to rare variants, three factors are at play. First, FWER can be elevated due to violations of the subset pivotality assumption in the built-in *maxT* multiple-testing correction procedure (Mahachie John et al. 2013). Second, when marker frequencies are rare, less than 10 individuals may contribute to a multi-locus genotype combination, in which case there is no power to assess whether this combination is related to a significantly higher or lower disease risk. As a consequence, the power to detect an interaction with such a combination may be hampered. Third, additive coding will always give rise to increased false positives, irrespective of whether rare or common variants are considered.

The fact that protocols #7 and #8 were the only ones that were able to highlight significant interactions, with either one of the 49 main effects SNPs listed in Evans et al. 2011, namely *rs2523608 x rs9788973* and *rs310787 x rs2844498* (Table 2), is not surprising. MB-MDR with additive encodings has a tendency towards generating more liberal test results than MB-MDR with co-dominant encodings (Mahachie 2012; Mahachie John et al. 2012). The SNPs *rs9788973* and *rs2523608* map to the genes *MAP2K4* and *HLA-B*. The *HLA-B* gene showed very strong association to AS (*rs4349859* p-value $<10^{-200}$) in (Evans et al. 2011) and was also related to AS in other studies (Jenisch et al. 1998; Nischwitz et al. 2010). In addition, the *rs2523608 x rs9788973* pair resides in the coding regions of the *HLA-B x MAP2K4* genes (Table 2), suggesting that AS pathology is not only linked to irregularities in peptide presentation to immune cells via major histocompatibility complex (MHC), but also to dysfunctions in intra-cellular signaling pathways.

Focusing on the common SNP pairs between GWAI protocols in our study (1230 pairs), only 3 showed a significant interaction in at least one protocol (Table 1), pointing towards the genes *MAGI3* and *PARK2*. The gene *MAGI3* controls intracellular signaling cell-cell adhesion and communication (Adamsky et al. 2003). In the context of AS, *MAGI3* potentially regulates cell-cell communication and adhesion of the cells in the inflamed joint areas between spinal discs and vertebra. *PARK2* was suggested before as a candidate gene for AS in (Claushuis et al. 2012). Mutations in the *PARK2* gene can cause alteration in cellular trafficking and protein degradation (Verdecia et al. 2003). In (Boisgerault et al. 1998), alterations in correct antigenic peptide presentation by major histocompatibility complex (MHC) class I molecules to $CD8^+$ T lymphocytes were linked with an early onset of chronic inflammation and AS. Further alteration in protein degradation, partially controlled by *PARK2*, may also suggest an alteration in the proper disposal of antigens. The aberrations in this process may potentially contribute to chronic inflammation of the spine followed by AS onset.

Only 20 pairs were common between our 8 protocols and the list of the 102 SNP x SNP interactions investigated in (Evans et al 2011). Clearly, several interesting pairs are missed by only looking at SNP pairs that are tested by all considered protocols (i.e. common SNP pairs). Imputation, to make the SNP x SNP pool more alike between protocols, may not only over-rule removal of SNPs after biofiltering (for which one may have had good reasons), it may also induce additional LD between SNPs, which may hugely increase false positives, depending on the analytic

tool used. Interestingly, 8 significant SNP x SNP interactions were detected for which at least one SNP was present in the 102 SNP pairs of (Evans et al. 2011). These 8 pairs involved the SNPs *rs30187, rs10050860* and *rs10781500,* and allowed to reproduce the statistically interacting gene pair *ERAP1 x HLA-B* reported in (Evans et al. 2011) via the interactions *rs3018 x rs2523608*, *rs10050860 x rs2523608* and *rs30187 x rs2523608* (Table 3). Notably, these findings were obtained with the only protocols using an additive main effects encodings (protocols #7 and #8); Evans and colleagues also primarily based their interaction testing on additive encodings.

However, by allowing more SNPs for interaction testing than in (Evans et al. 2011), we identified gene pairs not previously associated to AS: *ERAP1 x MICB*, *MICB x SNAPC4* and *HLA-B x SNAPC4* (Table 3), pointing towards interacting loci or regions between chromosome 5 and 6, and between 6 and 9. *MICB* is MHC Class I Mic-B Antigen linked to cell immune response and is functionally similar to MHC Class I encoded by the *HLA-B* gene. *MICB* is implicated in rheumatoid arthritis (Lopez-Arbesu et al. 2007). *SNAPC4* encodes small Nuclear RNA Activating Complex important for proper functioning of RNA Polymerase II and III. *ERAP1* encodes for endoplasmic reticulum aminopeptidase that trims peptides.

One of the top 480 common GO terms across GWAI protocol #1-#8 was GO:0002479 (Table S6). This term is functionally related to antigen processing and exogenous antigen presentation via MHC class I, TAP-dependent It may suggest that that AS pathology is partially caused by the inability of *ERAP1* amino-peptidase to correctly trim *HLA* class I-binding peptides and subsequently to present them to MHC complexes (Alvarez-Navarro and Lopez de Castro 2013). This possibly causes deregulation of the innate immunity and chronic inflammation of spine tissues that are typical symptoms displayed by AS patients (Chaudhary et al. 2011). Also appearing in the top 10 are GO terms linked to neural transmission processes (Table 4). This agrees with AS known disease pathology characterized by consistent pain and inflammation in the spine – part of the central nervous system (CNS). In particular, the GO terms highlighted in bold in Table 1 and Table S6 (column 1), even though based on the top 1000 SNP x SNP interactions (not necessarily statistically significant) may suggest a link between AS and mutations in genes involved in nerve impulse transmission and propagation (GO:0007411, GO:0007268, etc.). Furthermore, GO:0007219 (Table S6), linked to genes of the *Notch* signaling pathway (e.g. *RBP-J*, *PSEN1*, *ADAM10*), suggests AS interference with the correct development and growth of nerve tissue (Housden et al. 2013). It was shown by (Gao et al. 2013) that the *Notch* pathway also controls angiogenesis and that Vascular Endothelial Growth Factor (*VEGF*) and Angiopoietin (*Ang*) are both over-expressed in synovial tissues of Psoriatic Arthritis and Rheumatoid Arthritis patients.

Conclusions

Any GWAI analysis involves making choices about the input data (e.g., filtering using candidate genes or using prior biological knowledge), about LD-pruning thresholds, about adjusting for lower order effects (and how to encode these), and about the selection of the analytical tool (e.g., non-parametric, semi-parametric or fully parametric), as well as, the corrective method for multiple testing (Gusareva and Van Steen 2014). We have shown that even slight differences in protocols to perform a Genome-Wide Association Interaction (GWAI) study may hamper the results

reproducibility. We did so by applying the 8 GWAI protocols to real-life genome-wide SNP data on Ankylosing Spondylitis (AS) and controls.

Choices about marker selection (for instance filtering based on prior knowledge) are the most severe, as it may give rise to a dramatic reduction in SNPs for further GWAI analysis (Gusareva and Van Steen 2014; Sun et al. 2014; Van Steen 2012). Although biofiltering may reduce the ability to generate novel hypotheses about interactions (Sun et al. 2014), when doing so the effects of LD pruning and other protocol parameters seem to be less impactful on the final analysis results. More work is needed though to fully understand the interplay between LD-pruning and filtering strategies commonly adopted in GWAIs and to derive operational guidelines. In general, the second largest cause for heterogeneity in GWAI results is the adopted encoding scheme to adjust the interaction analysis for the lower-order effects (Gusareva and Van Steen 2014). The third largest cause is the adopted LD-pruning strategy. To date, no published work exists that comprehensively investigates the effect of LD on epistasis findings derived from several analytic tools. In order not to waste carefully acquired data, researchers are often tempted to adopt exhaustive screening tools whenever computationally feasible. As suggested in (Gusareva and Van Steen 2014), we nevertheless advocate LD-pruning at an r^2 of 0.75, to increase power, yet to reduce the generation of redundant (significant) SNP x SNP interactions. Exhaustively applying BOOST to LD-pruned AS data at an r^2 of 0.75 generated over 2,000 significantly interacting SNP pairs. The existence of moderate LD may induce multicollinearity in regression models and may increase the number of false positives (even when using a conservative multiple testing correction method such as Bonferroni). It shows that when applying a GWAI protocol, the results should be interpreted and discussed under the appropriate context, which includes the limitations and strengths of the adopted protocol, hereby addressing its different components.

Finally, with so many tools for GWAI analysis around, truly comparing these remains a challenging task in the absence of reference synthetic data sets that are rich enough to capture real-life complexities. Care has to be taken when “replicating” interactions with analytic tools that have a tendency to generate false positives: Can one be sure that one is not replicating a false positive? Clearly, no single tool will fit all. Tools are heterogeneous in their ability to recognize specific active epistasis modes and several such modes are likely to occur throughout the genome. This observation puts limitations to strategies that use agreement between different GWAI approaches as evidence for an interaction. It also favors the development of a hybrid SNP x SNP interaction detection tool, combining the best of several worlds when screening the genome.

Acknowledgments The research was funded by the Fonds de la Recherche Scientifique (FNRS) (incl. FNRS F.R.F.C. project convention n° 2.4609.11). We thank François Van Lishout and Elena Gusareva from the Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium for their support and advice. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award n° 076113.

Conflict of interest The authors declare that they have no competing interests.

References

Ackermann M, Strimmer K (2009) A general modular framework for gene set enrichment analysis. BMC Bioinformatics 10: 47. doi: 10.1186/1471-2105-10-47

Adamsky K, Arnold K, Sabanay H, Peles E (2003) Junctional protein MAGI-3 interacts with receptor tyrosine phosphatase beta (RPTP beta) and tyrosine-phosphorylated proteins. J Cell Sci 116: 1279-89.

Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22: 1600-7. doi: 10.1093/bioinformatics/btl140

Alvarez-Navarro C, Lopez de Castro JA (2013) ERAP1 structure, function and pathogenetic role in ankylosing spondylitis and other MHC-associated diseases. Mol Immunol. doi: 10.1016/j.molimm.2013.06.012

Boisgerault F, Mounier J, Tieng V, Stolzenberg MC, Khalil-Daher I, Schmid M, Sansonetti P, Charron D, Toubert A (1998) Alteration of HLA-B27 peptide presentation after infection of transfected murine L cells by Shigella flexneri. Infect Immun 66: 4484-90.

Bozeman M (2015) Golden Helix, Inc. SNP & Variation Suite (Version 7. x)[Software].

Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput: 368-79.

Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K (2011) Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. Ann Hum Genet 75: 78-89. doi: 10.1111/j.1469-1809.2010.00604.x

Chaudhary SB, Hullinger H, Vives MJ (2011) Management of acute spinal fractures in ankylosing spondylitis. ISRN Rheumatol 2011: 150484. doi: 10.5402/2011/150484

Claushuis D, Cortes A, Bradbury LA, Martin TM, Rosenbaum JT, Reveille JD, Wordsworth P, Pointon J, Evans D, Leo P, Mukhopadhyay P, Brown MA A genomewide association study of anterior uveiti Annual Scientific Meeting of the American-College-of-Rheumatology (ACR) and Association-of-Rheumatology-Health-Professionals (ARHP), Washington, DC, United States 2012. John Wiley & Sons, pp S259-S259

Colin Freeman JM (2012) GTOOL. Oxford University. http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html. Accessed March 2014

Dean LE, Jones GT, MacDonald AG, Downham C, Sturrock RD, Macfarlane GJ (2014) Global prevalence of ankylosing spondylitis. Rheumatology 53: 650-657.

Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, Oppermann U, Dilthey A, Pirinen M, Stone MA, Appleton L, Moutsianas L, Leslie S, Wordsworth T, Kenna TJ, Karaderi T, Thomas GP, Ward MM, Weisman MH, Farrar C, Bradbury LA, Danoy P, Inman RD, Maksymowych W, Gladman D, Rahman P, Spondyloarthritis Research Consortium of C, Morgan A, Marzo-Ortega H, Bowness P, Gaffney K, Gaston JS, Smith M, Bruges-Armas J, Couto AR, Sorrentino R, Paladini F, Ferreira MA, Xu H, Liu Y, Jiang L, Lopez-Larrea C, Diaz-Pena R, Lopez-Vazquez A, Zayats T, Band G, Bellenguez C, Blackburn H, Blackwell JM, Bramon E, Bumpstead SJ, Casas JP, Corvin A, Craddock N, Deloukas P, Dronov S, Duncanson A, Edkins S, Freeman C, Gillman M, Gray E, Gwilliam R, Hammond N, Hunt SE, Jankowski J, Jayakumar A, Langford C, Liddle J, Markus HS, Mathew CG, McCann OT, McCarthy MI, Palmer CN, Peltonen L, Plomin R, Potter SC, Rautanen A, Ravindrarajah R, Ricketts M, Samani N, Sawcer SJ, Strange A, Trembath RC, Viswanathan AC, Waller M, Weston P, Whittaker P, Widaa S, Wood NW, McVean G, Reveille JD, Wordsworth BP, Brown MA, Donnelly P, Australo-Anglo-American Spondyloarthritis C, Wellcome Trust Case Control C (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet 43: 761-7. doi: 10.1038/ng.873

Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ (2010) SCAN: SNP and copy number annotation. Bioinformatics 26: 259-62. doi: 10.1093/bioinformatics/btp644

Gao W, Sweeney C, Walsh C, Rooney P, McCormick J, Veale DJ, Fearon U (2013) Notch signalling pathways mediate synovial angiogenesis in response to vascular endothelial growth factor and angiopoietin 2. Ann Rheum Dis 72: 1080-8. doi: 10.1136/annrheumdis-2012-201978

Grange L (2014) Thesis: epistasis in genetic susceptibility to infectious diseases: comparison and development of methods application to severe dengue in Asia, Paris 7

Gusareva ES, Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet. doi: 10.1007/s00439-014-1480-y

Gyenesei A, Moody J, Semple CA, Haley CS, Wei WH (2012) High-throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics 28: 1957-64. doi: 10.1093/bioinformatics/bts304

Housden BE, Fu AQ, Krejci A, Bernard F, Fischer B, Tavaré S, Russell S, Bray SJ (2013) Transcriptional dynamics elicited by a short pulse of notch activation involves feed-forward regulation by E (spl)/Hes genes. PLoS genetics 9: e1003162.

Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13. doi: 10.1093/nar/gkn923

Jenisch S, Henseler T, Nair RP, Guo SW, Westphal E, Stuart P, Kronke M, Voorhees JJ, Christophers E, Elder JT (1998) Linkage analysis of human leukocyte antigen (HLA) markers in familial psoriasis: strong disequilibrium effects provide evidence for a major determinant in the HLA-B/-C region. Am J Hum Genet 63: 191-9. doi: 10.1086/301899

Kestler HA, Muller A, Gress TM, Buchholz M (2005) Generalized Venn diagrams: a new method of visualizing complex genetic set relations. Bioinformatics 21: 1592-5. doi: 10.1093/bioinformatics/bti169

Lopez-Arbesu R, Ballina-Garcia FJ, Alperi-Lopez M, Lopez-Soto A, Rodriguez-Rodero S, Martinez-Borra J, Lopez-Vazquez A, Fernandez-Morera JL, Riestra-Noriega JL, Queiro-Silva R, Quinones-Lombrana A, Lopez-Larrea C, Gonzalez S (2007) MHC class I chain-related gene B (MICB) is associated with rheumatoid arthritis susceptibility. Rheumatology (Oxford) 46: 426-30. doi: 10.1093/rheumatology/kel331

Mahachie J (2012) Thesis: Genomic Association Screening Methodology for High-Dimensional and Complex Data Structures, University of Liege

Mahachie John JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, Van Steen K (2011a) Comparison of genetic association strategies in the presence of rare alleles. BMC Proc 5 Suppl 9: S32. doi: 1753-6561-5-S9-S32 [pii]

10.1186/1753-6561-5-S9-S32

Mahachie John JM, Cattaert T, Lishout FV, Gusareva ES, Steen KV (2012) Lower-order effects adjustment in quantitative traits model-based multifactor dimensionality reduction. PLoS One 7: e29594. doi: 10.1371/journal.pone.0029594

Mahachie John JM, Van Lishout F, Gusareva ES, Van Steen K (2013) A robustness study of parametric and non-parametric tests in model-based multifactor dimensionality reduction for epistasis detection. BioData Min 6: 9. doi: 10.1186/1756-0381-6-9

Mahachie John JM, Van Lishout F, Van Steen K (2011b) Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. Eur J Hum Genet 19: 696-703. doi: 10.1038/ejhg.2011.17

Nischwitz S, Cepok S, Kroner A, Wolf C, Knop M, Muller-Sarnowski F, Pfister H, Roeske D, Rieckmann P, Hemmer B, Ising M, Uhr M, Bettecken T, Holsboer F, Muller-Myhsok B, Weber F (2010) Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. J Neuroimmunol 227: 162-6. doi: 10.1016/j.jneuroim.2010.06.003

Pang X, Wang Z, Yap JS, Wang J, Zhu J, Bo W, Lv Y, Xu F, Zhou T, Peng S, Shen D, Wu R (2013) A statistical procedure to map high-order epistasis for complex traits. Brief Bioinform 14: 302-14. doi: 10.1093/bib/bbs027

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-75. doi: 10.1086/519795

RCoreTeam (2013) R: A Language and Environment for Statistical Computing}. Vienna, Austria.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-47. doi: 10.1086/321276

Sun X, Lu Q, Mukheerjee S, Crane PK, Elston R, Ritchie MD (2014) Analysis pipeline for the epistasis search–statistical versus biological filtering. Frontiers in genetics 5.

Tabangin ME, Woo JG, Martin LJ (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc 3 Suppl 7: S41. doi: 10.1186/1753-6561-3-S7-S41

Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Theatre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K (2013) An efficient algorithm to perform multiple testing in epistasis screening. BMC Bioinformatics 14: 138. doi: 10.1186/1471-2105-14-138

Van Steen K (2012) Travelling the world of gene-gene interactions. Brief Bioinform 13: 1-19. doi: 10.1093/bib/bbr012

Verdecia MA, Joazeiro CA, Wells NJ, Ferrer JL, Bowman ME, Hunter T, Noel JP (2003) Conformational flexibility underlies ubiquitin ligation mediated by the WWP1 HECT domain E3 ligase. Mol Cell 11: 249-59.

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87: 325-40. doi: 10.1016/j.ajhg.2010.07.021

Wei W-H, Hemani G, Haley CS (2014a) Detecting epistasis in human complex traits. Nature Reviews Genetics.

Wei WH, Hemani G, Haley CS (2014b) Detecting epistasis in human complex traits. Nat Rev Genet. doi: 10.1038/nrg3747

Westfall PH (1993) Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons

Zhang X, Zou F, Wang W Fastanova: an efficient algorithm for genome-wide association study Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining 2008. ACM, pp 821-829

Zhang Y, Jiang B, Zhu J, Liu JS (2011) Bayesian models for detecting epistatic interactions from genetic data. Ann Hum Genet 75: 183-93. doi: 10.1111/j.1469-1809.2010.00621.x

Figures

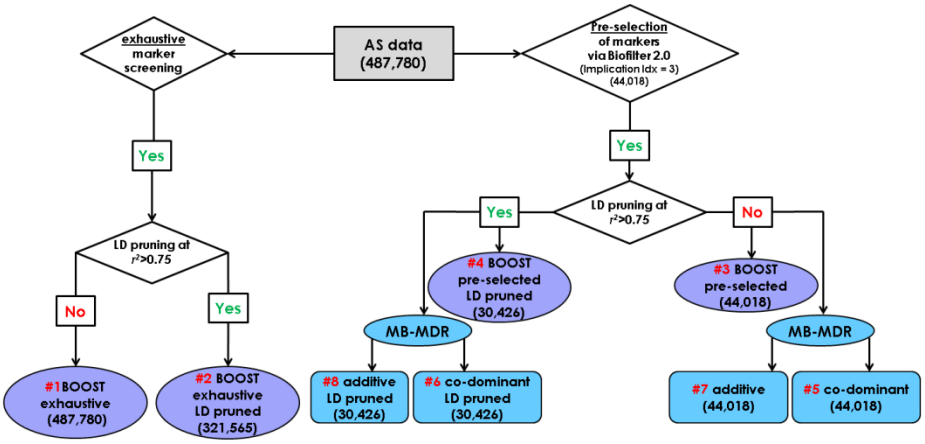
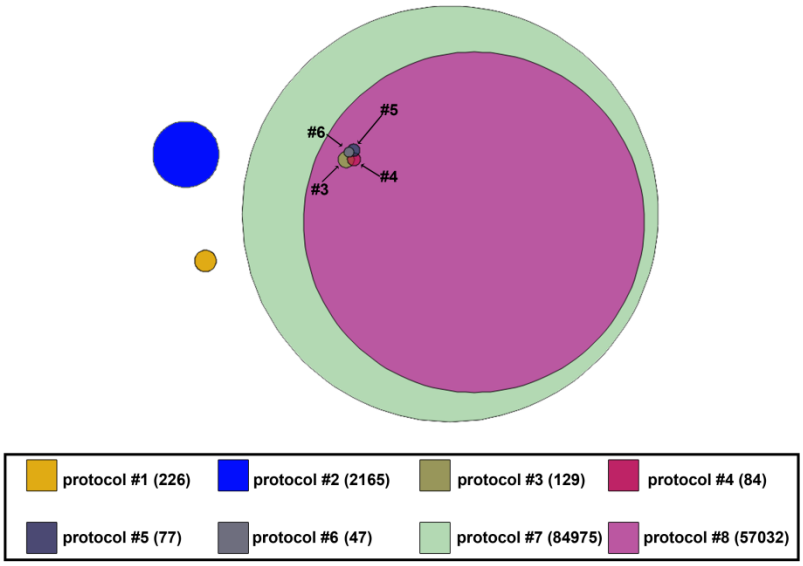


Fig. 1 Summary of 8 GWAI protocols included in this study and applied to AS data, the ankylosing spondylitis dataset from (Evans et al. 2011). The number of SNPs retained at each step is shown in parenthesis. The bottom nodes refer to GWAI protocol abbreviations and chosen parameters, following protocol components as described in (Gusareva and Van Steen 2014) GWAI protocol. The abbreviations additive and co-dominant refer to SNP main effects correction encodings in MB-MDR (see (Mahachie John et al. 2012)).

578
579



580

581 **Fig. 2** Euler diagram capturing significant SNP pairs identified in each of the 8 GWAI protocols. Each circle
582 represents a set of the significant SNP pairs in the corresponding GWAI protocol. Protocol numbers match the
583 protocol referencing used in Fig. 1.

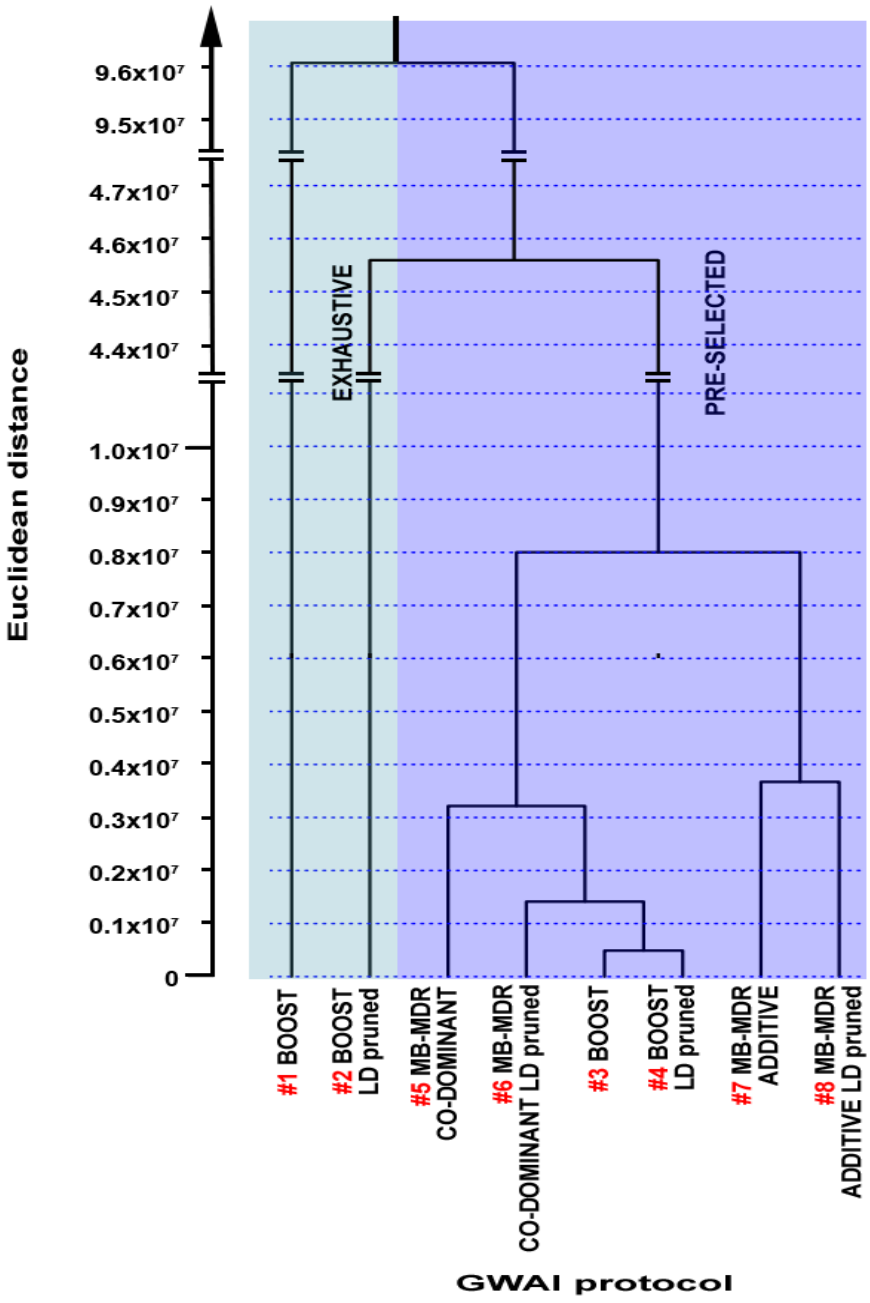


Fig. 3 Consistency between GWAI protocols based on 1230 common SNPs. Each SNP pair has a protocol-specific rank, which is stored in a protocol-specific vector. The dendrogram shows the distance between protocols, obtained via hierarchical clustering of 8 vectors (referring to the 8 GWAI protocols included in this study) of length 1230 and the Euclidean distance measure. The Euclidean distances themselves are listed in Table S2.

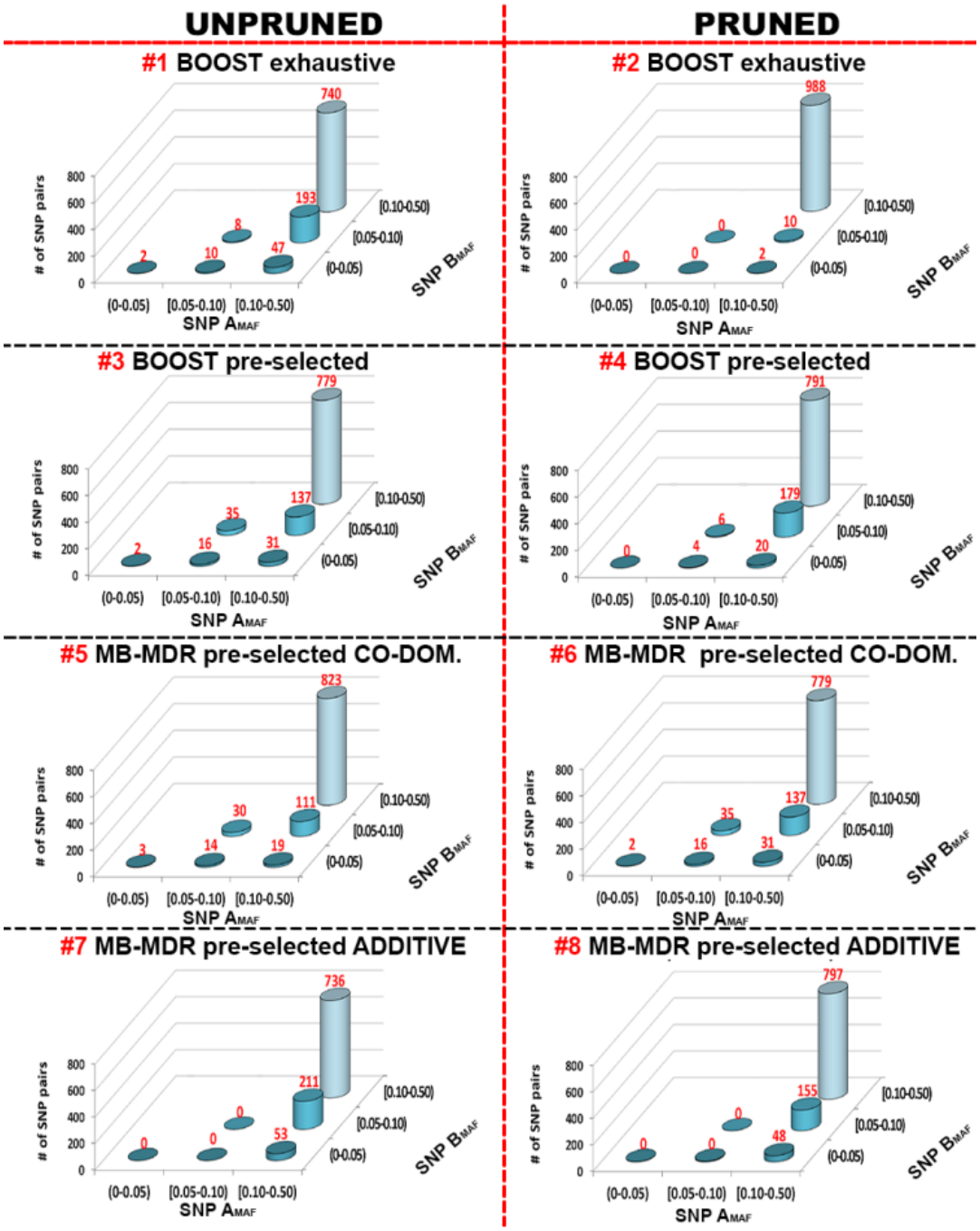


Fig. 4 Effect of SNP MAFs on ranked epistasis results. For each protocol, the top 1000 epistasis results are presented. Each SNP pair was ordered such that the SNP with the largest MAF was assigned to locus A, and the SNP with the lowest MAF to locus B. The numbers in red refer to the # of SNP pairs that were assigned to each 2-dimensional MAF bin.

598 **Tables**

599 **Table 1** – Most significant SNP pairs (among 1230 pairs) across 8 adopted GWAI analysis protocols. All *p*-values
600 are multiple testing corrected, either Bonferroni-based (BOOST protocols) or re-sampling based (MB-MDR protocols).
601

SNP A	SNP B	GWAI protocols								Gene A	Gene B
		BOOST				MB-MDR					
		#1	#2	#3	#4	#5	#6	#7	#8		
rs12026423	rs7528311[+]	0.009	0.004	7.72E-05	3.69E-05	0.401	1	0.001	0.004	*MAGI3*	*MAGI3*
rs11964796	rs13194019[++]	1	1	0.024	0.012	0.401	1	1	0.995	*PARK2*	*PARK2*
rs13194019	rs1784607[+++]	1	1	0.144	0.069	0.401	1	1	0.995	*PARK2*	*PARK2*

602 [+] rs12026423/rs7528311 are separated by 13833 bp., $r^2 = 0.0178$; [++] rs11964796/ rs13194019 are separated by 9824 bp. and $r^2 =$
603 0.0309; [+++] rs13194019/rs1784607 are separated by 3127 bp. and $r^2 = 0.0610$

604

605 **Table 2** Significant pairs containing one of the 49 SNPs associated to main effects (Evans et al. 2011), obtained via
606 the 8 GWAI protocols.

SNP A	SNP B	GWAI protocols								Gene A	Gene B
		#1	#2	#3	#4	#5	#6	#7	#8		
		multiple testing adjusted p-values									
rs2523608	**rs9788973***	1	1	1	1	1	1	0.001	0.001	*HLA-B*	*MAP2K4*
rs30187*	rs2844498	1	1	1	1	1	1	0.001	0.002	*ERAP1*	*NA*

607 *SNPs that occurring as main effects SNPs in Supplementary Table 2 of (Evans et al. 2011) are highlighted in **bold**.
608

609

Table 3 – Statistically significant SNP x SNP interactions that contain a SNP occurring in at least one of 102 SNP pairs listed in Supplementary Table 5 in **Evans et al. 2011***.

GWAI protocol	SNP A	SNP B	Chr A	Chr B	*p*-value	Gene A	Gene B
#8	**rs30187***	rs2844498	5	6	0.002	*ERAP1*	*MICB*
	rs30187*	rs2523608	5	6	0.038	*ERAP1*	*HLA-B*
#7	**rs10050860***	rs2844498	5	6	0.001	*ERAP1*	*MICB*
	rs10050860*	rs2523608	5	6	0.001	*ERAP1*	*HLA-B*
	rs30187*	rs2844498	5	6	0.001	*ERAP1*	*MICB*
	rs30187*	rs2523608	5	6	0.001	*ERAP1*	*HLA-B*
	rs2523608	**rs10781500***	6	9	0.001	*HLA-B*	*SNAPC4*
	rs2844498	**rs10781500***	6	9	0.001	*MICB*	*SNAPC4*

* - SNPs that were analyzed in Supplementary Table 5 by (Evans et al. 2011) are highlighted.

Table 4 Top 10 Significant GO terms related to top 1000 SNP pairs per GWAI protocol, based on Fisher's combined *p*-value at a significance level of 0.05. Protocol-specific p-values are also reported.

GO ID	GO Term Description	GWAI protocols								combined*
		#1	#2	#3	#4	#5	#6	#7	#8	
GO:0007411	**axon guidance**	**5.18E-02**	**1**	**4.00E-16**	**4.40E-18**	**1.90E-12**	**2.20E-15**	**1.20E-13**	**5.70E-16**	**7.86E-77**
GO:0030168	platelet activation	5.83E-01	1	2.90E-15	2.30E-15	3.20E-11	1.20E-10	4.10E-09	1.20E-11	3.95E-58
GO:0055085	**transmembrane transport**	**4.74E-02**	**1.55E-01**	**1.80E-09**	**1.00E-09**	**3.20E-11**	**5.40E-11**	**6.00E-09**	**1.00E-12**	**3.04E-50**
GO:0007268	**synaptic transmission**	**2.17E-02**	**1**	**8.00E-10**	**3.10E-08**	**1.50E-06**	**2.40E-09**	**6.30E-07**	**5.00E-08**	**2.00E-36**
GO:0007173	epidermal growth factor receptor signaling pathway	2.10E-02	1	7.80E-10	1.40E-11	2.40E-07	6.80E-07	2.40E-05	7.20E-06	1.55E-34
GO:0008543	fibroblast growth factor receptor signaling pathway	9.85E-02	1	5.40E-08	6.90E-11	5.10E-07	1.80E-08	2.20E-04	3.60E-04	2.99E-30
GO:0007202	activation of phospholipase C activity	1.03E-02	1	2.60E-08	9.40E-09	1.80E-06	6.80E-06	5.10E-06	3.90E-06	6.44E-30
GO:0006112	energy reserve metabolic process	1.76E-01	1	9.90E-07	3.40E-09	1.20E-04	1.80E-07	5.90E-06	3.60E-05	1.46E-26
GO:0042493	response to drug	1.31E-01	5.62E-01	2.70E-05	1.40E-09	5.06E-03	9.80E-05	1.90E-07	6.60E-08	7.90E-26
GO:0006198	cAMP catabolic process	5.17E-03	1	5.10E-04	2.50E-05	2.90E-06	5.60E-08	1.00E-05	1.50E-06	6.04E-25

* - Combined p-values summarize information across the 8 considered protocols. The most relevant GO terms for AS are indicated in **bold**, as well as, GWAI-specific *p*-values when < 0.05. The exhaustive list of significant GO terms is shown in Table S6.