King's Research Portal

DOI:
10.1109/CIBCB.2013.6595394

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Ibrahim, Z. M., Newhouse, S., & Dobson, R. (2013). Detecting epistasis in the presence of linkage disequilibrium: A focused comparison. In *The International IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 96-103). IEEE. https://doi.org/10.1109/CIBCB.2013.6595394

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 01. Dec. 2021

Detecting Epistasis in the Presence of Linkage Disequilibrium: A Focused Comparison

Zina M. Ibrahim, Stephen Newhouse and Richard Dobson
King's College London
De Crespigny Park, London SE5 8AF
United Kingdom
{*zina.ibrahim,stephen.newhouse,richard.j.dobson*}*@kcl.ac.uk*

***Abstract*—We present results from a comparison of three epistasis-detection tools using large-scale simulated genetic data: SNPHarvester, SNPRuler and Ambience. The tools were chosen based on their merits to be representative of the state of the art of epistasis detection. We design and conduct experiments to test the performance of the methods in detecting interacting loci or their proxies in linkage disequilibrium (LD) tagged regions, in datasets containing simulated 2,3 and 4-way epistatic interactions.**

The results show that SNPHarvester is the fastest while Ambience is the most robust. Moreover, SNPRuler provides the best power, specially with higher-level interactions, but cannot scale-up to larger datasets.

***Keywords*-Genome-wide association data, SNP-SNP Interactions, Epistasis, Linkage Disequilibrium.**

I. INTRODUCTION

It is largely thought that many human complex traits and disorders (e.g. Height, Diabetes and Dementia) are caused by the joint effects of multiple genetic variations [1]. The most common type of DNA sequence variations are single-nucleotide polymorphism (SNPs), which occur when a single nucleotide in the genome takes different forms or alleles between paired chromosomes in an individual [2]. Studying human genetic variation has successfully been exploited by genome-wide association studies (GWAS) to explore the associations between SNPs and phenotypes [6], in order to help unveil new disease mechanisms and develop better strategies for detection, treatment and prevention of disease [7]. Many analyses, however, focus on single SNPs or small combinations of SNPs, testing for association SNP by SNP.

In complex traits, genetic variants responsible for any phenotype may show little or no marginal association with the phenotype, when considered individually, but display a strong joint interaction effect. These interactions are known as *epistatic interactions* [3], whose discovery and characterization can provide great insight on complex traits and has been shown to be a defining factor in the genetic causes of complex diseases such as breast cancer [4] and Alzheimer's disease [5].

From a computational standpoint, using GWAS data to detect epistasis introduces several fundamental challenges that prohibit the use of classical statistical and machine learning algorithms. These problems stem from the nature of the data being extremely high-dimensional (a typical analysis involves over half a million variants and thousands of samples), offering much fewer samples than variables and containing many non-linear interactions that cannot be identified using standard statistical models.

Consequently, different efforts to adapt machine learning models to scale up to the complexity of the problem have been made. The resulting models include genetic programming [8], Random Forests [9], neural networks [10], Bayesian Networks [11], [12], [13], greedy search [14], information theoretical approaches [15] and Multifactor Dimensionality Reduction (MDR) [16].

The literature contains many reviews of SNP-SNP interaction models [17] [1], [18], [19], [2], [20], [21], [22]. These reviews pinpoint the strengths and weaknesses of existing methods with respect to 1) ability to detect interactions when no main effects are present 2) computational efficiency 3) the quality of the detected interactions 4) the ability to deal with higher order interactions (i.e. the number of interacting SNPs responsible for the trait in question).

Despite their differences, all the reviews have the common theme of specifically looking for known or simulated disease SNPs (or their combinations). Because genotyping platforms were designed to tag common genetic variation and without prior knowledge of the true disease causing variants, it is likely that any classifier will miss the true causal loci, unless by luck, the genotyped SNPs contain the true causal SNP(s). The problem decreases when making SNPs that are in *linkage disequilibrium* (LD) [23], [24] with the causal SNP to be the focus of the study, in conjunction with the SNP itself. LD is defined as the non-random association among the genetic components in the genome. SNPs that lie in the same LD block or region, are highly correlated and tend to be inherited together. Therefore, when including LD SNPs in the search for interactions, the chance of picking up SNP neighbourhoods that contain the causal variant increases. The candidate *blocks* can then be used to perform a more focused analysis.

This work aims at complementing the current studies by investigating how epistasis-detection tools perform in detecting SNPs that are in LD with the target causal SNPs. More specifically, we test the hypothesis of whether promising tools can detect linkage disequilibrium using large-scale genomic data in a true epistasis environment. This work is motivated by the vision of implementing an epistasis-detection pipeline which uses LD for dimensionality reduction as outlined earlier.

In addition to this focus, we also study how the tools perform when the number of interacting SNPs is greater than 2 (in addition to the highly-studied 2-way interactions case) in which no main effects are present. For the sake of correctness, we do not use real datasets but instead simulate high-dimensional data that mimics real GWAS data in which epistasis has been embedded.

In our experiments, we study the performance of SNPRuler [25], SNPHarvester [14] and Ambience [15]. Both SNPRuler and SNPHarvester have reported a better performance compared to other tools as they 1) can handle the absence of main effects [26], 2) have been shown to handle 3-way cases [2], [15] and 3) have been stated to deal with LD (although this is to be tested in our work).

We feel that Ambience has been understudied by the recent reviews. It has only been tested for the two-way case where it has been shown to perform as well as logistic regression and outperform MDR [22]. [26] excludes AMBIENCE from its review because BOOST has been shown to outperform logistic regression [27] for the two-way case. Because we are interested in higher-order interactions, we choose not to use these reports as basis for discarding it. This is especially so given that AMBIENCE has been designed for performance with higher-order interactions (as Section III-B details), so we include it in our work.

We exclude Random Forests [9] and Bayesian Models (BEAM and PBEAM) [28], [12] because we are interested in tools that have been reported to detect interactions in the absence of main effects. Both random forests and BEAM require main effects to perform well. BOOST [27] has reported good performance in the absence of main effects but will not be included in our study as it has only been designed for 2-way interactions. Finally, we exclude statistical models that can only deal with 2-way interactions because they have been shown to be unable to scale up.

The paper is structured as follows. We start by introducing epistasis and linkage disequilibrium from a computation point of view in Section II and the tools used in this study in Section III. In Section IV, we detail the method we followed in simulating our data and describe the settings under which the experiments are run. The experimental results are given in Section V, followed by a summary in Section VI.

II. DETECTING EPISTASIS: AN OVERVIEW

A. *Problem Formulation*

The problem of finding n-way interactions from GWAS data can be formulated as follows:

Given:

- N individuals.
- A set S_M of M SNPs (features).
- A matrix X of $M \times N$ variables, one per SNP per individual. $\forall m, n$, $1 \leq n \leq N$, $1 \leq m \leq M$ X_{mn} is a discrete categorical variable, describing the value of SNP m for individual n.
- Phenotype Y manifested in N phenotype variables, one per individual $Y_1, ..., Y_N$, such that $1 \leq n \leq N$,Y_n is binary and always known.

Output: K Subsets $S_1, ..., S_K \subset S_M$ such that elements of each set possess a epistatic interaction with respect to the phenotype Y.

B. *Computational Challenges*

1) Curse of dimensionality: which recognizes that the problem's search space grows exponentially with the number of dimensions (SNPs). This is complicated by the fact that the number of interacting SNPs is mostly unknown, and most of the studies search for 2-SNP interactions within the data. The problem becomes much complex when we are searching for more than 3-way interactions which are characteristic of complex genetic diseases. In these cases, the multiple interacting vectors is not necessarily navigable by heuristic algorithms such as a greedy search.

2) $p >> n$ problem: GWAS studies have fewer examples (n) than features (p). This is because a large number of SNPs (up to one million) are genotyped from a small number (a few thousands) of biological samples (subjects): a severe case of the $p >> n$ problem. For example, in the Wellcome Trust Case Control 2 Study (WTCCC2), there were typically only 5000 individuals for datasets of 500,000 SNPs. This problem increases with SNP imputation, where millions of variants can be tested for association.

3) Non-linearity of epistatic interactions: makes performing single-SNP statistical filtering tests (e.g. chi-2 test, Fisher's exact test [1] to select variants based on some threshold a dangerous approach. This is because a group of SNPs may have an interaction effect without displaying any marginal association with the phenotype when considered individually [1], [3]. This problem can be generalized to any k-way tests: not displaying a k-way interaction does not necessarily preclude higher order interactions. Hence, feature space reduction through these types of approach are generally not recommended. This is further complicated by the fact that what SNPs do functionally, or which SNPs interact is mostly unknown. Consequently, exhaustive search to cover all possible two-way, three way, ..., N-way interactions maybe required, resulting a task which is computationally impossible to achieve [1].

C. Linkage Disequilibrium (LD)

Is defined as the non-random association of alleles at multiple SNPs and occurs when alleles at two loci are inherited together. The two loci are therefore tightly correlated and their frequencies exhibit a non-random association. The existence of LD prohibits the use of standard Machine Learning approaches which assume independence among the variables (e.g. multivariate regression). Such approaches do not take into account the interactions due to LD, resulting in complexities that are difficult to deal with due to the high dimensionality of the data [3].

For two loci with genotypes *Aa* and *Bb*, if the two loci are in LD, then the corresponding SNPs (e.g *A* and *B*) will show dependence, i.e. $p_{AB} \neq p_A p_B$. In this regard, an intuitive measure of LD is the *disequilibrium coefficient*, which uses the difference between the two probabilities $D_{AB} = p_{AB} - p_A p_B$. However, the value of D_{AB} depends on the allele frequencies, it would be difficult to compare D_{AB} between markers, rendering the measure hard to interpret. Because of this, a more preferred measure is r^2, the square of the correlation coefficient between the two alleles at the corresponding loci [23], [24], computed as follows:

$$\begin{aligned} r^2 &= \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} \\ &= \frac{\chi^2}{2n} \end{aligned}$$

Where p_A and p_B are the allele frequencies of A and B respectively and SD_{AB} is the disequilibrium coefficient. Essentially, $0 \leq r^2 \leq 1$. A value of 1 indicates that the markers are always inherited together while a value of 0 indicates independence.

III. THE TOOLS

A. Greedy Search: SNPHarvester

SNPHarverster [14] is a method to detect interactions that exhibit weak marginal effects in GWAS [2]. It is a stochastic search method which selects a set of candidate SNP groups from hundreds of thousands of SNPs efficiently.

Given a dataset of N SNPs, S_N, SNPHarvester operates as follows: First, all SNPs which show main effects are removed from the dataset. The second step consists of a heuristic search algorithm called *PathSeeker* which aims at identifying M-way interacting SNPs for a fixed M. The algorithm works by initially selecting a random set of M SNPs: $S_M = \{s_1, s_2, ..., s_M\}$ and initializing a statistical score r_M (e.g. z-statistics) for the set. Then the algorithm swaps each member of S_M with each of the remaining SNPs (S_N / S_M) and records whether or not the new set increases the score r_M and the new score if it shows statistical significance. When the run ends, extracting the optimal set of SNPs S_M, the entire step is repeated with this set being the initial set fed to the algorithm.

The usefulness of SNPHarvester stems from its efficiency in reducing the pool of SNPs to a manageable size so that existing tools for epistatis detection (e.g. MDR) can be directly applied to the SNP groups it outputs.

B. Information Theoretic Models: Ambience

Is an entropy-based system, which formulates a novel and parsimonious measure of information gain (IG), phenotype-associated information (PAI). PAI is specifically designed so that for a set of variables, it does not require contributions from PAI values of lower-order combinations of the variables. This property makes it suitable for detecting SNP-SNP interactions as it does not result in an exponential number of computations for higher-order combinations.

For a set of N SNPs $X_1, ...X_N$ and phenotype variable P, $PAI(X1, ...X_n, P)$ is calculated as in Equation 1 below:

$$PAI(X_1, .., X_N, P) = TCI(X_1, ..X_N, P) - TCI(X_1, ..X_N) \quad (1)$$

Where TCI is the amount of information shared among the variables in the set [15] and is defined in terms of the entropies of the individual variables and the entropy of the joint distribution [15] as in Equation 2 below:

$$TCI(X_1, X_2, ..., X_K) = \sum_{i=1}^{N} H(X_i) - H(X_1, ..., X_N) \quad (2)$$

When TCI is zero, the variables are essentially independent and knowledge about one gives no information about the others, while a higher values indicate more information being provided about the rest of the variables from one of them. Hence, TCI is a general measure of dependency.

As Equation 1 shows, PAI is obtained from the TCI measures of the genetic variables and the phenotype by removing the interdependencies among the genetic variables. As a result, the measure is robust with respect to LD correlations (among other confounding factors) [15]. This, along with the efficiency of computation inherent to the measure makes it a good choice for ranking interactions.

The Ambience algorithm is a greedy-search algorithm which essentially receives as input a set of SNP variables, a phenotype variable and parameters ρ and θ denoting the order of interactions and the number of iterations respectively. The algorithm iteratively computes the PAI values of the 1-way, 2-way, ..., θ-way, retaining the top θ combinations in terms of their PAI values at each step and using them to start the next higher-order iteration. The combinations output by Ambience designate regions in the combinatorial space with highly-interacting variables.

Ambience has been shown superior power compared to MDR in two-locus studies [22], especially with low heritability. As far as we know, our study is the first to compare Ambience with other tools for higher-order interactions.

C. Rule-based Approaches: SNPRuler

The idea behind SNPRuler is that each epistatic interaction induces a set of rules [25] describing the relationships between feature (SNPs) and class (phenotype) variables. For instance, given two loci with genotypes *Aa* and *Bb*,,then one simple rules which can be induced is: *if SNP1 has genotype AA, and regardless of the genotype of SNP2, then the probability of this sample to be a disease sample is 0.66.*

SNPRuler detects epistatic interactions by modelling them as rules and then apply a predictive rule-learning algorithm. The efficiency of the algorithm stems from the hypothesis that learning rules is much easier and faster than searching for and evaluating interactions. This is because not all rules embedded in the data are necessary to deduce the underlying interactions; if one is identified, a fast validation step can be used to find the others [25].

Rules are built using tree expansion by starting from single SNPs and building different branches designating different possible rules. Each rule is evaluated at every step for adequacy using mutual information $I(S_i, ..., S_j)|\varsigma$ as given in Equation III-C below:

$$I(S_i, ..., S_j|\varsigma) = \sum p(s_i, ..., s_j, \varsigma) log \frac{p(s_i, ...s_j|\varsigma)}{p(s_i|\varsigma)...p(s_j|\varsigma)} \tag{3}$$

Where ς designates the data and $S_i, ..., S_j$ desginates all the SNPs involved in the tree branch (and constituting the current rule) starting with SNP S_i and ending with SNP S_j. A rule continues to be built as long as its corresponding MI does not exceed a previously-decided upper-bound. Once the tree is constructed, every path from the root to a leaf constitutes a rule embedding allele-level epistatic interactions. The possible interactions induced by each rule are then evaluated using depth-first search and ranked based on their utility, which is computed using χ^2. The interactions can then be pruned and ordered by their corresponding utility and given as output.

IV. EXPERIMENTAL SETTINGS

A. Data Simulation

We generated large data sets that show epistatic gene-gene interaction effects without any main effect in order to compare the performance of the above methods on different scales. We simulated case-control data sets using reference haplotypes from the HapMap Phase 3 project (August 2009 CEU haplotypes) - NCBI Build 36 (dbSNP b126)[*] since it targets SNPs with minor allele frequencies (MAF) $> 5\%$. We selected chromosome 22 as a pool for the haplotypes to be used for the simulation. The chromosome has a total of 119,317 SNPS, making it adequate for testing our models.

From the pool of SNPs, we used HAPGEN2 [†] to generate 12 different data sets using combinations of 10,000, 30,000, 60,000 and 100,000 SNPs and 1,000, 5000 and 10,000 samples. Unlike other studies (e.g. [2]), we do not simulate datasets of 1000 SNPs as our aim is to test the application of methods for a range of large datasets.

Since HAPGEN2 cannot generate epistatic interactions, we first used it to generate data under the null hypothesis (no disease effect) resulting a total of 36 data sets based on the sample sizes and number of SNPs as defined above. The interacting disease SNPs were identified as follows. Firstly the datasets were LD pruned to make sure that the interacting SNPs did not belong to the same LD block; this way, the sets of SNPs they are in LD with will be more or less disjoint. We then selected the disease SNPs from the haplotype such that they all have other SNPs in LD when setting R^2 = 0.8 and maximum genomic distance = 500bp.

We then simulated interactions using an accompanying R package called `simulateDiscretePhenotypes` [‡]. The package implements three models of interactions, out of which we chose the multiplicative effect between and within loci detailed in table I for the two-SNP case. In the table, α corresponds to the baseline odds ratio of having the disease and (1 + θ) is the increase in the odds given when one of the two disease alleles (a or b) are present. Consequently, a person with genotype Aa or Bb has odds $\alpha(1+\theta)$ of getting the disease while a person with genotype aa or BB has odds $\alpha(1+\theta)^2$ of getting the disease. The model multiplies the odds if either loci contains a disease allele.

Since `simulateDiscretePhenotypes` only implements the 2-SNP case, we first extended its functions to accommodate 3-way and 4-way interactions. The odds of getting the disease for the n-way case is given below as (odds$_n$).

odds$_n$ = $\alpha \times (1+\theta_1)^{(a_1)} \times ... \times \alpha \times (1+\theta_n)^{(a_n)}$

For each data set, we use alleles with Minor Allele Frequency (MAF) above 0.05 and set α 0.5 in order to obtain comparable sizes for case and control sets.

Table I
MODEL 1: TWO-LOCUS MULTIPLICATIVE DISEASE EFFECT BETWEEN AND WITHIN LOCI

	AA	**Aa**	**aa**
BB	a	a(1+θ)	a(1+θ)^2
Bb	a(1+θ)	a(1+θ)^2	a(1+θ)^3
bb	a(1+θ)^2	a(1+θ)^3	a(1+θ)^4

B. Hardware

We ran the simulation on a cluster with HP C7000 enclosures, each containing 15 x HP BL460c G6 blades. Each blade is configured with 2 x 4-core Intel Xeon Processor X5550. Of the 30 blades, 15 blades contain 78 GB RAM and the remaining 15 blades contain 54 GB of RAM.

---

[*] http://mathgen.stats.ox.ac.uk/impute/impute_v2_retired6Aug2010.html#Download

[†] https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html

[‡] https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html#Interaction

V. EXPERIMENTS AND RESULTS

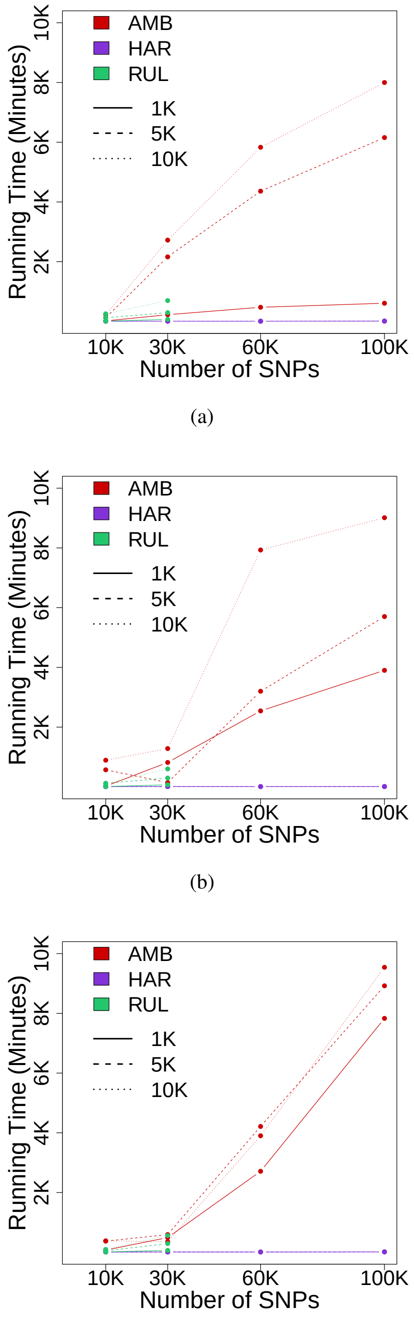
A. *Run Time (Scalability)*

The runtime results are shown in Figures 1(a)-(c) for datasets embedding 2-way, 3-way and 4-way interactions for each sample size and number of SNPs. In each figure, the horizontal axes designate the number of SNPs for the setting (10,000, 30,000, 60,000 and 100,000), while the vertical axes record the running time in minutes. The results are color-coded to distinguish the time taken by each tool and are drawn with different line types to distinguish the runs for the different sample sizes as the legends show.

The results show that SNPHarvester (coloured purple) is the fastest under all settings, with runtime not exceeding two hours and not varying greatly when using larger sample sizes and increasing number of SNPs. This is why the three lines representing SNPHarvester in each figure occlude, appearing as one solid line. SNPRuler (coloured green) collapsed when increasing the number of SNPs to 60,000 and, reporting errors in heap-allocation, explaining the lack of lines for the 60,000 and 100,000 SNPs settings. The speed of SNPRuler matches that of SNPHarvester for the 10,000-SNP case but becomes slower with 30,000 SNPs.

Ambience is the slowest among the three algorithms, with its runtime reaching seven days for the 100,000-SNP case. Moreover, the effect of increasing the number of samples is more visible in the case of Ambience, with days separating the runtime between datasets having 1,000 samples and those having 10,000 samples.

All three tools behave as expected, taking longer to run with more interacting SNPs. Due to the speed of SNPHarvester however, this effect is not as visible as in the other two tools. In Ambience on the other hand, it can be seen that the tool takes longer to run in the 4-way case (Figure 1(c)) than it does for the two other cases, with runtime increasing almost monotonically as the number of samples and interacting SNPs are increased. This is especially so for datasets with smaller samples sizes (i.e. the 1,000 sample case represented by a solid red line in each figure).



(a)

(b)

(c)

Figure 1. Running Time for the Ambience, SNPHarvester and SNPRuler for a) 2-way b) 3-way and c) 4-way interactions.

B. *Detection of LD SNPs*

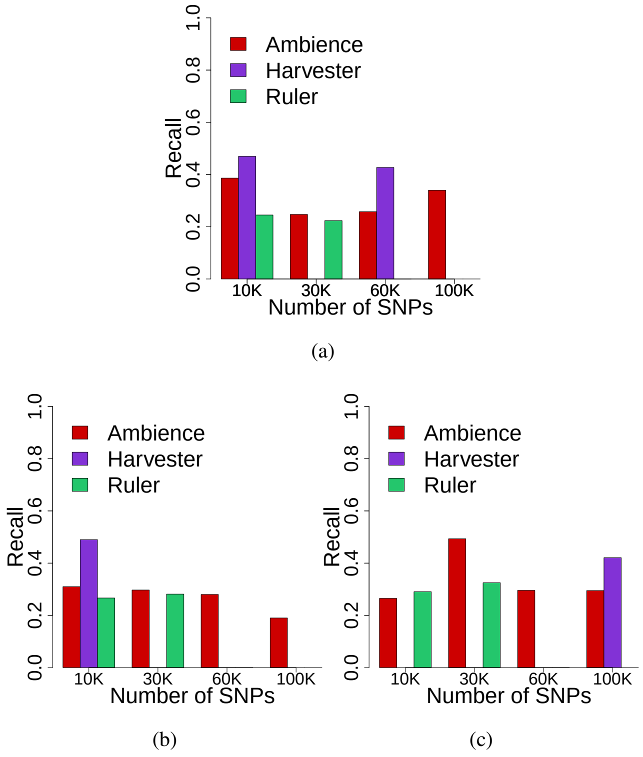
When collecting the output given by the three tools, we found that none of the tools successfully identified the right combination of SNPs in its result sets. We therefore based our conclusions on whether the tools identify at least one of the target SNPs (disease or LD SNP) in its output. These findings are reported in Figures 2(a)-7(c). Figures 2(a)-4(c) present recall of each method for the 2-way, 3-way and 4-way interactions, while 5(a)-7(c) show their corresponding precision values respectively. In all the figures, the results concerning each method are color-coded as before (Ambience: red, SNPHarvester: purple, SNPRuler: green) and are grouped by SNP size (the horizontal axes). Moreover, each figure is divided into three subfigures, reporting results for the 1K, 5K and 10K sample sizes.

SNPHarvester failed to identify any SNP more often than not (only produced some true positives in 16 of the 36 runs), making it the most unreliable of the three tools. However, when harvester does produce output, it seems to almost always provide more true positives than the other tools, making both its precision and recall relatively higher.

Ambience seems to be the most consistent of the three tools, only failing to produce true hits once in the 3-way case (Figures 3(c), 6(c)) and four times in the 4-way case (Figures 4 and 7). It produces better precision and recall values for the 2-way case (Figures 2 and 5). However, as the order of interactions increases, its performance tends to fall behind.

While SNPRuler did not run for the 60K and 100KSNP cases, its recall and precision rates for the 10K and 30K was slightly less than that of Ambience in the 2-way case (Figures 2 and 5) and exceeded Ambience's for higher-interactions (Figures 3 - 4 and 6 - 7). Moreover, it only failed to output any LD SNPs in the 4-way, 30K SNPs and 10K samples case (Figure 4(c)).

Overall, both precision and recall decrease as the number of interacting SNPs increases for all the tools. In addition, recall and precision values seem to fluctuate for the three tools across the different runs, but are more or less comparable, indicating more or less similar true positive to false negative and true positive to false positive ratios respectively (Figures 2-4 to 5-7).

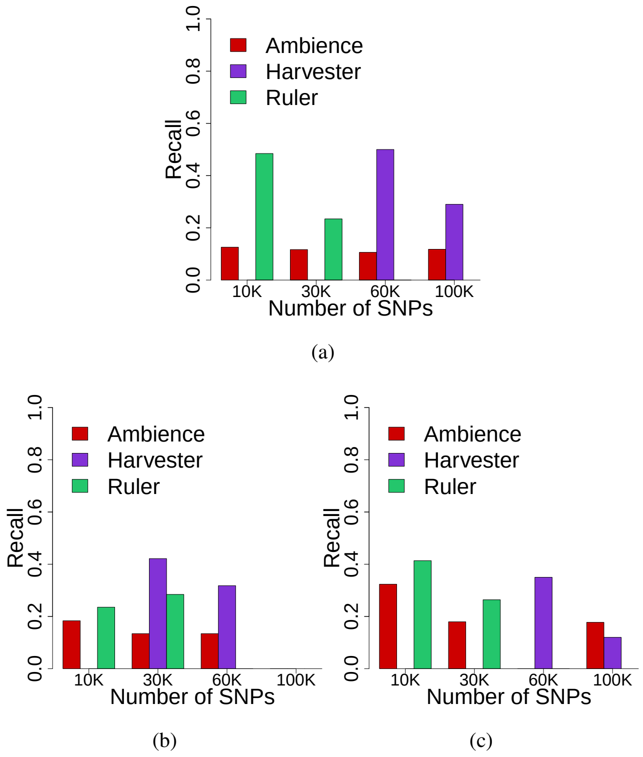


(a)

(b)

(c)

Figure 2. Recall for 2-way Case with a) 1K sample size b) 5K sample size c) 10K sample size



(a)

(b)

(c)

Figure 3. Recall for 3-way Case with a) 1K sample size b) 5K sample size c) 10K sample size

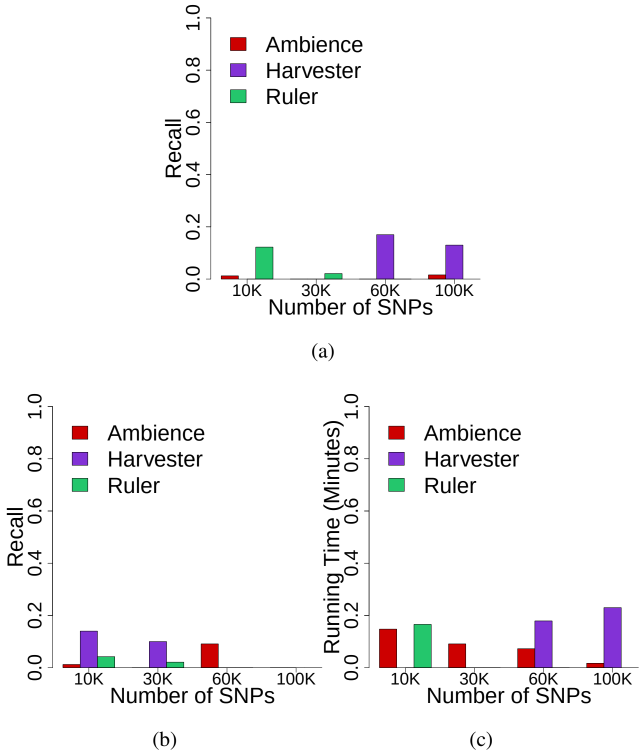
VI. CONCLUSIONS AND FUTURE WORK

We presented a study to evaluate the performance of Ambience, SNPHarvester and SNPRuler in detecting LD SNPs in simulated large-scale GWAS data using 36 different settings. The data embeds 2-way, 3-way and 4-way epistatic interactions showing no main effects and has sizes that range between 10K and 100K SNPs and 1K and 10K samples.

Overall, and contrary to prior reports, no method was able to detect interactions in a true epistasis settings. Moreover, no method is superior in all aspects. SNPHarvester is extremely fast but has less power than the other two tools. While Ambience is capable of detecting target SNPs at a better rate than the other tools, it misses more true SNPs than the other tools, reporting lower precision and recall rates as as the order of interactions increases. Finally, SNPRuler has good power compared to Ambience and is definitely more stable than SNPHarvester, but cannot scale up to the size of real GWAS datasets.

REFERENCES

[1] H. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Genome-Wide Association Studies,* vol. 10, pp. 392-404 , 2009.



(a)

(b)

(c)

Figure 4. Recall for 4-way Case with a) 1K sample size b) 5K sample size c) 10K sample size

[2] L. Chen, G. Yu, D. Miller, L. Song and C. Langefeld, "A ground truth based comparative study on detectin epistatic SNPs," *Proceedings of the IEEE Conference on Bioinformatics and Biomedicine,* 2009, pp. 26-31.

[3] H. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in human," *Human Molecular Genetics* 1;11(20) 2463-2463 (2002).

[4] R. Scully, "Epistatic Relationships in the BRCA1-BRCA2 Pathway," *PLoS Genetics,* vol. 7, no. 7, pp. e1002183, 2011.

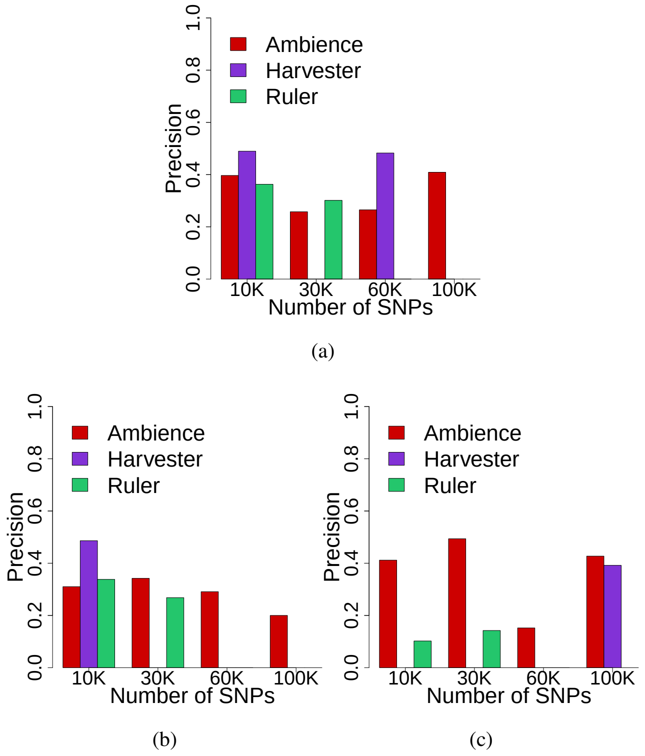
[5] O. Combarros, M. Cortina-Borja, A. Smith and D. Lehmann, "Epistasis in sporadic Alzheimer's disease," *Neurobiology of Aging,* vol. 30, no. 9, pp. 1333-1349, 2009.

[6] A.Johnson and C. O'Donnell, "An Open Access Database of Genome-wide Association Results," *BMC Medical Genetics,* vol. 10, no. 6, 2009.

[7] T. Manolio, "Genomewide association studies and assessment of the risk of disease," *New England Journal of Medicine,* vol. 363, no. 2, pp. 166-176, 2010.

[8] M. Iglesias, V. Penaranda, C. Vidal and A. Verschoren, "Higher Epistasis in Genetic Algorithms," *Bulletin of the Australian Mathematical Society,* vol. 77, no. 2, pp. 225-243, 2008.

[9] D. Schwarz, I. Knig and A. Ziegler, "On Safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," *Bioinformatics,* vol. 26, no. 14, pp. 1752-1758, 2011.



(a)

(b)

(c)

Figure 5. Precision for 2-way Case with a) 1K sample size b) 5K sample size c) 10K sample size

[10] S. Turner, S. Dudek and M. Ritchie, "Grammatical evolution of neural networks for discovering epistasis among quantitative trait loci," *Proceedings of the 8th European conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics,* 2010, 86-97..

[11] X. Jiang, R. Neapolitan, M. Barmada and S. Visweswaran1, "Learning Genetic Epistasis Using Bayesian Network Scoring Criteria," *BMC Bioinformatics,* vol. 1, pp. 89-91, 2011.

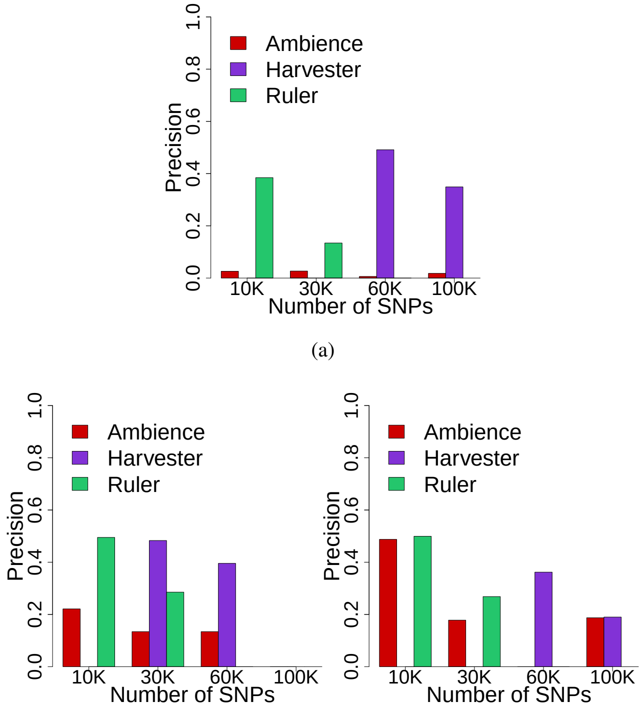
[12] T. Peng, P. Du and Y. Li, "PBEAM: A parallel implementation of BEAM for genome-wide inference of epistatic interactions," *Bioinformation,* vol. 3, no. 8, pp. 349351, 2009.

[13] Y. Guan and M. Stephens, "Bayesian variable selection regression for genome-wide association studies and other large-scale problems," *Annals of Applied Statistics,* vol. 5, pp. 1780-1815, 2011.

[14] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue and W. Yu, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics,* vol. 25, no. 4, pp. 504-11, 2009.

[15] P. Chanda et al, "AMBIENCE: A Novel Approach and Efficient Algorithm for Identifying Informative Genetic and Environmental Associations With Complex Phenotypes," *Genetics,* vol. 180, no. 2, pp. 1191-1210, 2008.

[16] M. Ritchie et al, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism



(a)

(b)

(c)

Figure 6. Precision for 3-way Case with a) 1K sample size b) 5K sample size c) 10K sample size

genes in sporadic breast cancer," *American Journal of Human Genetics,* vol. 69, pp. 138-147, 2001.

[17] B. McKinney, D. Reif, M. Ritchie and J. Moore, "Machine Learning for Detecting Gene-Gene Interactions," *Journal of Applied Bioinformatics,* vol. 5, no. 2, pp. 77-88, 2006.

[18] S. Szmyczak et al, "Machine Learning in Genome-Wide Association Studies," *Genetic Epidemiology* vol. 33, pp. S51-S57, 2009.

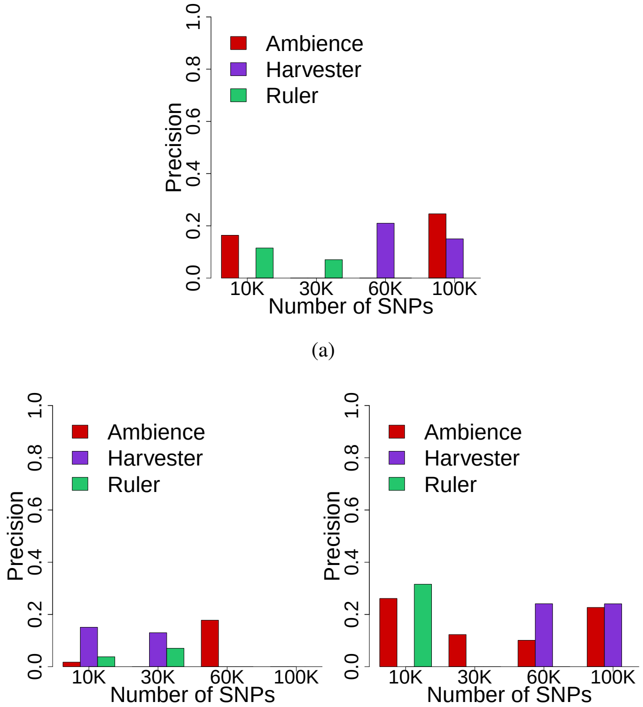
[19] X. Zhang, S. Huang, F. Zou and W. Wang, "Tools for Efficient Epistasis Detection in Genome-Wide Association Studies," *Source Code for Biology and Medicine,* vol. 6, pp. 1-4, 2011.

[20] A. Motsinger-Reif, D. Reif, T. Fanelli and M. Ritchie, "A Comparison of Analytical Methods for Genetic Association Studies," *Genetic Epidemiology,* vol. 32, pp. 767-778, 2008.

[21] K. Van Steen, "Travelling the World of Gene-gene Interactions," *Briefings in Bioinformatics* vol. 13, no. 1, pp. 1-19, 2012.

[22] L. Sucheston, P. Chanda, Z. Zhang, D. Tritchler and M. Ramanathan, " Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity," *BMC Genomics* vol, 11, pp. 487, 2011.

[23] M. Slatkin, "Linkeage disequilibrium–understanding the evolutionary past and mapping the medical future," *Nature Reviews in Genetics,* vol. 9, no. 6, pp. 477-485, 2008.



(a)

(b)

(c)

Figure 7. Precision for 4-way Case with a) 1K sample size b) 5K sample size c) 10K sample size

[24] J.Pritchard and M. Przeworski, "Linkage disequilibrium in humans: models and data," *American Journal of Human Genetics,* vol. 69, pp. 1-14, 2001.

[25] Xiang Wan et al, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics,* vol. 26, no. 1, pp. 30-37, 2010. 26

[26] Yue Wang et al, "An empirical comparison of several recent epistatic interaction detection methods," *Bioinformatics,* vol. 27, no. 21, pp. 2936-2943, 2011.

[27] X. Wan et al, "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *American Journal of Human Genetics,* vol. 87, no. 3, pp. 325-340, 2010.

[28] Y. Zhang and J. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics,* vol. 39, pp. 1167-1173, 2007.

[29] S. Purcell et al, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics,* vol. 81, no. 3, pp. 559-575, 2007.

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.