A Novel Approach to Detecting Epistasis using Random Sampling Regularisation

Jade Hind, *Member, IEEE*, Paulo Lisboa, *Senior Member, IEEE*, Abir J. Hussain, *Member, IEEE*, Dhiya Al-Jumeily, *Senior Member, IEEE*

Abstract— *Epistasis is a progressive approach that complements the 'common disease, common variant' hypothesis that highlights the potential for connected networks of genetic variants collaborating to produce a phenotypic expression. Epistasis is commonly performed as a pairwise or limitless-arity capacity that considers variant networks as either variant vs variant or as high order interactions. This type of analysis extends the number of tests that were previously performed in a standard approach such as* Genome-Wide Association Study (*GWAS), in which* False Discovery Rate *(FDR) is already an issue, therefore by multiplying the number of tests up to a factorial rate also increases the issue of FDR. Further to this, epistasis introduces its own limitations of computational complexity and intensity that are generated based on the analysis performed; to consider the most intense approach, a multivariate analysis introduces a time complexity of* O(n!). *Proposed in this paper is a novel methodology for the detection of epistasis using interpretable methods and best practice to outline interactions through filtering processes. Using a process of Random Sampling Regularisation which randomly splits and produces sample sets to conduct a voting system to regularise the significance and reliability of biological markers, SNPs. Preliminary results are promising, outlining a concise detection of interactions. Results for the detection of epistasis, in the classification of breast cancer patients, indicated eight outlined risk candidate interactions from five variants and a singular candidate variant with high protective association.*

Index Terms—GWAS study, SNPs, Artificial Intelligence, genome, logistic regression.

1. Introduction

Breast cancer is a complex disease; multifactorial effects represent the phenotypic response of the subject. While there is currently an abundance of techniques for the analysis of genetic data, there is still a limited contribution of reproducible genetic signals that provide evidence of association with sporadic breast cancer. This study aims to investigate the interactions that exist in subjects associated to breast neoplasms in invasive breast cancer. By considering a representative set of Single nucleotide polymorphisms (SNPs) from the genome, further analysis can be conducted from genome-wide analysis to suggest potential SNPs for interactions.

Current efforts in breast cancer have led to early screening, with great successes in reducing the number of advanced cases. Further to this, the introduction of genetic knowledge also outlines patient and their family for potential susceptibility to cancer through examples of the BRCA1 and BRCA2 gene. These measures provide a preventative outlook for patient health, a leading direction to personalised medicine that will address patients on an individual basis taking into consideration factors such as genetic make-up. The identification of these SNPs could lead to the classification of susceptible breast cancer patients and potentially the pharmacological or therapy treatments that are most suitable for these individuals.

Progressive approach epistasis, invites new avenues of research. The phenomenon that suggests combinations of biological material such as SNP variants are working as a system or network to produce the phenotypic outcome is becoming a favorable lead. Given the elusiveness of genetic causation in the face of high heritability, epistasis suggests an enigmatic genetic component is not being detected as the signal for networks of SNPs are masking one another. Therefore the research subject of epistasis invites a new problem area to explore. Current practices in epistasis detection range from pairwise to exhaustive search criteria; the limiting nature of pairwise detection could lead to loss of information by oversight however exhaustive search present their own problems with the demand for computational power.

To address the current issue of computational complexity, our method proposes a multidimensional reduction technique that relies on the strong presence of SNPs (Single Nucleotide Polymorphisms) across multiple panels of subjects with the objective being to outline prominent SNPs with unyielding presence among the panels. Considering the current limitations of epistasis, a programme that employs a limitless-arity approach is utilised to identify relationships between SNPs. Further analysis investigates the true significance of these associations, adopting a variety of robust statistical techniques that overcome issues that are common in predictive algorithms, such as overfitting. Described in this paper are the problems further defined and the chosen solutions to address these issues within the proposed methodology.

- *Jade Hind with LivingLens Limited Company, 49 Jamaica St, Liverpool, UK, Email: jade.hind@livinglens.tv*
- *Paulo Lisboa with the Department of Applied Mathematics, Liverpool John Moores University, Byrom Street, Liverpool, L33AF, Email: P.G.lisboa@ljmu.ac.uk*
- *Abir. J. Hussain, and Dhiya Al-Jumeily, with the Department of Computer Science, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK. E-mail: {A.Hussain , D.AlJumeily}@ljmu.ac.uk.*

2. BACKGROUND AND RELATED WORKS

The analysis of genomic data is providing individuals with information and knowledge to take control of their health [1]; while still in its early stages, genomic analysis and its resulting outcomes can aid the healthcare sector, approaching the much debated subject of personalised medicine [1]. Personalised medicine caters for the needs of patients by considering their biological and epidemiological make-up. This will in future replace the current "one-size-fits-all" approach that is common in prescription medication; an individual's biological make-up could provide information for the most appropriate treatment response, i.e. indicating the amount, variety and response to particular drugs [2]. This is important in complex diseases as treatment is often based on a necessary 'trail-and-error' period which may or may not provide relief from the symptoms [3]. Further to this, even t reatment options that aid in reducing or eliminating the problematic symptoms of a disease or disorder, can also cause a variety of side effects that can still effect patients Quality of Life [4][7] [5].

There are several approaches to genomic study which are commonly used to identify risk variants in common, complex diseases such Breast Cancer; GWAS (Genome-wide Association study), Candidate Gene and Familial studies. One of the most popular genetic feature inputted for study analysis are SNPs (Single Nucleotide Polymorphisms), these are variants in base pairs within the DNA sequence [8] [6]. While a majority of these SNPs will have little to no impact on the biological systems, the consequential causal sequence can lead to imbalances in chemicals, misfolds in protein polypeptide chains and instability in mRNA transcripts [9] [11-15]. The involvement of these SNPs in the genetic analysis for the purpose of finding risk variants is due to the abundance of variation throughout the genome; proving promising and successful in many determined diseases so far [9][10][16]. Although there are many other genetic and biological studies that are successfully undertaken, the following identified approaches utilise the SNP feature input for the analysis of correlation and susceptibility in subjects. As such, the following sub-sections encompass the approaches to data analysis that consider the variability that exists in genomics due to the structure of genetic data and the subject specification.

2.1 Genome-Wide Association Studies

Genome-Wide Association Studies (GWAS) provides a way in which the whole genome (genotyped SNPs) can be scanned to identify SNPs that confer risk for the identified and analysed phenotype. Presenting a hypothesis-free approach that has introduced an option for researchers to visualise whole genome effects for diseases.

The most common approach in GWAS utilises a case-control setup [17]; Cases refer to a cohort affected by the disease subject of the study and Control refers to a cohort who are unaffected by the disease. The proceedings of a GWA study aims to find the correlation results between the cohorts and the disease. In an ordinary case-control GWAS, the odds ratio is the first considered statistics in which an $OR > 1$ suggests the association of an allele is a risk for disease, the greater the difference from 1, the more indicative of an association and an $OR < 1$ suggests a protective association against a disease [10]. Performing a chi-squared test from the results will provide significance of the alleles association; that is, how likely it is that the result is truly associated with the disease.

While GWAS presents a unique approach for analysis of genetic material, its requirements introduce both advantages and disadvantages. GWAS have also previously been acknowledged for their expense; however, this criticism is becoming obsolete as advances in technology are reducing the costly price [18]. This approach also outlines some disadvantages that effect the reliability of the study such including high false discovery rate and the overlooking of rare alleles which could potentially be important to the discovery of biomarkers [19]. As such, an important feature of GWAS are the requirements for a large sample size for reliability of result outcomes [28]. Unfortunately, this accommodation does not rectify the issues that are present with false discovery in GWAS and given the parameters that define the size of these studies, transfer learning is commonly adopted from methods that aim to reduce, rectify and eliminate the effects of bias and false discovery in 'Big Data'. A common approach from big data techniques is to use multiple testing adjustments, as discussed later on in Methodology. Successes in GWAS have previously outlined viable SNPs in complex diseases such as Crohn's Disease [29], Rheumatoid Arthritis [30-33] and Celiac Disease [34]. It has also previously been proposed that GWAS studies should be a first step in the genetic identification process [43].

Within the next section, the focus moves to an approach that contrasts with the whole-genome approach of GWAS to introduce an approach that focuses its efforts in areas of significance based on prior knowledge.

2.2 Breast Cancer

Cancer is a global concern, with prominent mortality rates across the board demonstrated by its current position as second leading cause of death in the United States, 2016 [45]. In 2016, 11,563 deaths were reported due to Breast Cancer, with increasing incidence rates that resulted in approx. 55,000 new cases in 2015, for England alone [48]. The current survival rate for Breast Cancer in England is 78%, however this is highly related to screening practices that are in place for quick diagnosis, ensuring treatment is started as soon as possible [48].

The symptoms of breast cancer vary, and quite often are due to common occurrences in the body that are unrelated to the development of cancerous cells. Current campaigns urge women to regularly check the size, shape and feel of breasts to be aware of changes that are associated with breast cancer. Lumps, breast pain, changes in skin colour and texture, abnormal discharge and inverted or sunken nipples encompass the most common symptoms associated with Breast Cancer [49].

Breast cancer is most curable in its early stages which emphasises the importance of the screening processes that are in place. Diagnosis of breast cancer is most commonly conducted using imaging techniques including mammograms and ultrasound [50]. Diagnosis of breast cancer normally adheres to a 'two-week wait' protocol that insists that suspected cancer patients are first seen by a specialist within 2-weeks [51]. With this protocol in place, ~90% of cases with known stage are diagnosed with early stage breast cancer (Stage 1 & 2, discussed later) [48].

Breast Cancer is divided in to 4 stages that are based upon the TNM staging system. The TNM system uses information about the tumour size, node spread and metastasis status to assign a

stage to a case. Genetic predisposition to breast cancer is fast becoming a common practice in aiding both the diagnostic and preventative measures for Breast Cancer[52], [53]. One of the most commonly associated but rare genetic associations in breast cancer is the BRCA1 and BRCA2 genes; these are inherited genes that express a predisposition to breast cancer in 15% of familial cases; presenting a 50-85% increased risk in women. BRCA1 an BRCA2 presents the highest penetrance in familial cases of breast cancer, however several genes have been indicated to present a percentage of penetrance for familial breast cancer but does not explain all [54].

Familial studies encompass the vast majority of successful genetic discoveries in breast cancer with emphasis being placed in the now well-known BRCA1 and BRCA2 genes. There are currently three established categories of mutations defined as high penetrance, moderate-risk and low-risk. These categories currently include a number of genes indicated in research including ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, FANCM, MLH1, MRE11A, MSH2, MSH6, MUTYH, NBN, PALB2, PMS1, PMS2, PTEN, RAD50, RAD51C, STK11 and TP53 genes [54] [55] [20-27].

TABLE I: PENETRANCE LEVEL OF ESTABLISHED SNPS ASSOCIATED WITH BREAST CANCER

High Penetrance	Gene	Incidence
	BRCA1	82% lifetime risk
	BRCA2	82% lifetime risk
	PTEN	85% lifetime risk
	TP53	25% by age 74
	CDH1	39% lifetime risk of lobular breast cancer
	STK11	32% by age 60
Moderate Risk	**Gene**	**Risk in Females (RR)[a]**
	CHEK2	1.7
	BRIP1	2.0
	ATM	2.37
	PALB2	2.3

[a] RR; Relative Risk

As a high penetrance mutation, the BRCA1 gene was first localised in 1990 by Hall et al. [60] who utilised logarithm of the likelihood ratio for linkage, or better known as 'Lod', to ascertain a likelihood ratio ranging from 2000:1 and 1.4×10^{6}:1 among the 23 tested families within the study [60]. From this, further studies were performed, leading to the discovering of the BRCA2 gene by Wooster et al. [61], using similar techniques. Table I provides approximate estimates for penetrance and relative risk of high and moderate penetrance SNPs [54], [56-59].

2.3 Epistatic Studies

Epistasis association is a developing technique that investigates the role of multiple genetic signatures in respect to the disease, suggesting the interacting components produce the phenotypic expression commonly associated with the disease. Breast cancer has received a lot of attention using the epistasis technique within the past 10 years. However, having been subject to the limitations of epistasis, studies have been focusing their effort on smaller sets of biologically related gene or prior knowledge from previous studies [62].

The limitations of this study type result in many epistasis studies focusing on a dramatically reduced set of SNPs or using a limited 2-way interaction model that only considers the interactions of 2 SNPs in relation to the phenotype. A large-scale analysis of ~89,000 subjects and 75,380 SNPs previously identified via 9 GWAS studies encompassing 10,052 cases and 12575 controls was conducted using two-way SNP interactions [63]. This study yielded few SNPs that exceeded the genome-wide threshold of 1×10^{-8} but concluded more SNPs with 1×10^{-6}. Further studies have been conducted in association with Breast Cancer, using reduction parameters for SNP dimension such as pathway analysis. Pathway analysis considers the pathology of the disease and uses these genetic signatures to conduct an epistasis study. Using DNA repair, modification and metabolism related pathways, Sapkota et al [64] identified 2-way SNP interactions that yielded a result of $<7.3\times10^{-3}$, however this again uses a two-way interaction model which may not confer the risk that is associated with a group of interacting SNPs across genes or chromosomes.

2.4 Data Description

Subject genotypes were attained from repository platform, Database of Genotypes and Phenotypes (DBGaP). Data was collected under the Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative that funded 5 projects, one of which was the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) project that focused its efforts in breast cancer for the systematic discovery and replication of additional common genetic variants. These variants were assessed for their biological significance and from this, developed evidence-based assessments of the clinical validity of prediction algorithms in practice.

Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and the National Cancer Institute. The following studies, used within this study, contributed germline DNA from breast cancer cases and controls:

- Copenhagen General Population Study (CGPS)
- Cancer Prevention Study 2 (CPSII)
- Melbourne Collaborative Cohort Study (MCCS)
- Multiethnic Cohort (MEC)
- Nurses Health Study (NHS)
- Nurses Health Study 2 (NHS2)
- Women of African Ancestry Breast Cancer Study (WAABCS)
- Women's Health Initiative (WHI)

With Breast Cancer incidence rates primarily effecting the female population, the study cohort is made up entirely of female participants (n = 28,281). Of these participants 14,435 subjects were cases and 13,846 were controls. Age ranged from 20 to 98 ($\mu = 63$) based on a sample of 27,585 with age ranging from 20 to 92 ($\mu = 65$). Cases were split into 3 histology types, invasive (12,412), in-situ (1,506) and unknown (517) of which, individuals of interest in this research are invasive histology type. Invasive breast cancer status regards cancer cells that have at least 'spread' to the surrounding breast tissue.

3. Techniques and Methods

3.1 Proposed Methodology

Random Sampling Regularisation is our proposed methodology that aims to improve selection criterion for epistasis, reduce false

discovery and cater for non-intensive computational requirements. The proposed method consists of 6 stages described as follows.

A. Stage 1: Quality Control (QC)

The proposed methodology adheres to standard practices in GWAS [66][67]; adopted in many studies [65][60] employing standard processes; ancestry divergence, sex inconsistencies, heterozygosity, relatedness and duplicates in subjects, Linkage Disequilibrium (LD) pruning and common threshold measures, Minor Allele Frequency (MAF), and Genotype rate (GENO). Conservative threshold measures are applied to create a reliable dataset that is devoid of missing values and information that could cause errors later [35-42].

Table II records the number of removed observations and features after each step has been performed and provides a breakdown of the processes. The remaining dataset is comprised of 13,649 (7,136 cases) (6,513 controls) observations and 320,247 features, or SNPs.

TABLE II: QUALITY CONTROL PROCESS EXCLUSION VALUES

Process	Removed		Remaining	
	Subjects	Variants	Subjects	Variants
Before QC			**28281**	**528620**
Ancestry Divergence	675	-	27606	-
Related and Duplicates	7750	-	19856	-
Heterozygosity	301	-	19555	-
Sex Inconsistencies	72	-	19483	-
Linkage Disequilibrium Pruning	-	116115	-	412505
Threshold Measures				
Missingness in Individuals	4399	-	15084	-
Genotype Call Rate	-	21561	-	390944
Hardy Weinberg Equilibrium	-	1269	-	389675
Minor Allele Frequency	-	69428	-	320247
Missing Phenotype	1355	-	13729	-
Exc. Crit: Age < 40	80	-	13649	-
After QC			**13649**	**320247**

B. Stage 2: Cohort Extraction

To respond to the common issues that are present in GWAS's, the cohort extraction is a preliminary stage in Random Sampling Regularisation to extract random cohorts of individuals that can represent a real-world sample cohort for analysis. Using this method, the data is prepared to explore in later stages if a constant effect is common across significant SNPs (which it should be if it is truly associated). Further to this, the effect of population structure is evident in many studies; this is a difficult problem to solve unless the study data has been obtained from a purpose built clinical study. Therefore, the randomisation of the data can disperse the effect of the population structure among the cohort to reduce the effects. Having performed QC, excluding any outlier subjects and/or SNPs, in a standard GWA study the remaining subjects are used to perform an association analysis to provide a resulting set of SNPs with their corresponding p-values to indicate the probability of significance to the phenotype.

C. Stage 3: Association Analysis

Association analysis models vary in the outcome information, this puts importance on choosing the most appropriate model for the approach . As further analysis is used to investigate the information beyond this point, there is affordance to use the allelic model analysis and gain the benefits of the increased statistical power. During this stage, multiple testing is not utilised but will be addressed in later stages, however genomic control is used to control for population structure.

D. Stage 4: Feature Selection

This stage uses the results of the association analysis to produce a subset of features that show significant association to the given phenotype. The results from the association analysis are combined to produce a mean gc-value and the corresponding standard deviation which will provide information as to how much the value is shifting across the subset cohorts. This will provide information as to whether the SNP is consistently associated with the phenotype or is falsely associated with a sample of subjects. Continuously mentioned in literature, is the 'statistical power' of a study, within genomic studies it is generally accepted that the bigger the cohort the less likely a SNP will show false associations; this is due to the normalisation of data with the addition of more observations. While this is true, consider that the number of features that are tested during genomic studies is large and as a result the likelihood of producing a false positive is also increased. By splitting the cohorts into n×n sections, our sample size is improved by n times.

E. Stage 5: Epistasis

The purpose of this stage is to sift through the combinations of SNPs to outline potentially significant relationships for further analysis. The use of a software programme that employs a limitless arity approach produces exhaustive results that investigate the relationships that exist between all combinations (excluding those eliminated during reduction techniques) and the focus phenotype. This benefits the methodology as it considers a larger feature set that would otherwise be impractical to explore via normal statistical techniques. We have chosen to use LAMPlink due to the benefits of limitless arity with the additional benefit of speed. Acknowledging the use of a dominant model leads to sacrificing potential combinations, the purpose of the method is to explore the effects of using random sampling regularisation to produce a set of resulting candidates while reducing FP and FN error rates.

LAMPlink provides a method of detecting significant associations using a large number of features. Generally epistasis programmes will perform epistatic interaction tests using two-way feature sets e.g. PLINK; this significantly reduces the exploratory power of epistasis by by-passing the potential for component clusters of 3 or more features. LAMPlink tests the potential of every possible combination while reducing the number of tests performed by adjusting the number of SNPs based on [4] complexity correction. This significantly reduces computational complexity and also reduces the time consuming process that is generally associated with epistasis approaches. The following section outlines the process of LAMPlink.

F. Stage 6: Inference Analysis

During this stage, the relationships outlined by LAMPlink are further analysed. As LAMPlink is only used to outline the potential relationship, this stage is used to expand the relationships outlined and to further analyse them to confirm or disregard the findings. Relationships that adhere to the petal plot policy will be extracted from the training set, with allelic and genotypic states combinations explored; as all combination states would be exhaustive to perform manually. During stage 6, interaction rela-

tionships outlined by LAMPLINK are further analysed. Combinations including singular, two-way and three-way along with all possible combination states (excluding combinations that do not contain more than 10 values in one cell of case-control based contingency table) are analysed to expand and explore the relationships outlined during the previous stage. In order to outline the most significant relationships, all combinations having successfully been analysed using the test dataset will be further analysed for their relationship to breast cancer in terms of penetrance, incidence and risk.

Penetrance (Pe): How many subjects have been affected by the phenotype that also carry the genomic interaction state?

$$P(Phenotype \mid Genotype) = \frac{a}{a+b}$$

Incidence(I): What percentage of the sample population carry this genomic interaction state?

$$P(Phenotype \mid SamplePopulation) = \frac{a+b}{n}$$

Risk (OR): How strongly associated is the presence of the genomic interaction state with the presence of the phenotypic state?

$$\frac{odds\ of\ disease\ among\ exposed}{odds\ of\ disease\ among\ unexposed} = \frac{ad}{bc}$$

These measures will provide a real-world understanding of how this information relates to breast cancer and what the resulting combinations indicate statistically.

3.2 Evaluation and Validation

Each aspect of this methodology has carefully considered and serves a purpose for filtering and extraction of features of significance approaching the problem for the detection of epistasis. The methodology aims to address a series of issues as shown in Table III.

TABLE III: BENEFITS OF THE PROPOSED METHODOLOGY.

Reducing FDR
Stage 2 introduces the first stage that aims to address the issue of FDR by tackling the problem before it occurs. Finalising in Stage 3 using multiple association analysis to build a picture of how genetic component behaves among varying sample populations.
Computational expense
Exhaustive Epistasis searches requires extended time allowances or require substantial hardware for processing. This method aims to outline a representative set of SNPs that do not require substantial hardware but is a small enough set of representative SNPs that epistasis can be performed on in a reasonable time constraint.
Improvement of epistasis detection
One of the most challenging problems in epistasis is the detection of interactions while accounting for the influencing pitfalls of FDR, computational complexity and the statistical filtering that is commonly used to reduce this. This method aim to outline the most prominent SNPs for epistasis from a large feature set that can commonly become lost in the expanse of information.
Concise identification of interaction combinations
Further to the identification of SNP for epistasis, the aim of this method is to concisely outline combinations that show significance with the phenotype. Commonly many combinations will be outlined for significance with the phenotype due to FDR and SNP selection; the aim of this method is to combat these issues.

4 Simulation Results

This section presents the results obtained using the proposed methodology as outlined in the previous Section 3.1. In stage 2, random cohort sampling is performed for our test data**. Stage 2** provides the outcome frequencies for the subsets of individuals that are to be used in the following stage.

For **Stage 3**, association analysis is performed using 2 different approaches; standard case-control and proposed random sampling regularisation method. Both approaches were conducted using the same techniques only varying the input information. An allelic model, adjusted by genomic control included remaining subjects and SNPs from the quality control stage with variation conducted for the Random Sampling Regularisation approach using cohort samples as reported in 'Random Cohort Sampling'.

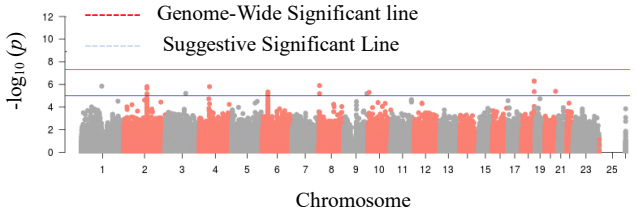

Figure 1: Manhattan plot for standard case-control method using allelic model

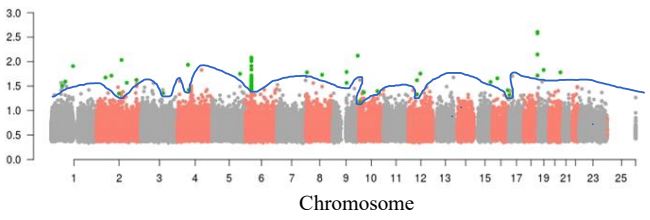

Figure 2: Manhattan Plot generated from mean of 9 SNP p-values using random sampling Regularisation method scaled for comparison to Standard case-control

Figure 1 visualises the p-values produced from an association analysis using standard case-control approach in a Manhattan plot, while Figure 6 visualises the p-values produced using the Random Sampling Regularisation approach. Visible is the clear decrease in significance for all SNPs. As the mean of 9 p-values for each SNP is used to create the mean p-values, any sample p-values that show little significance for the SNP will reduce the mean p-value but it will reduce the presence of False Positives based on chance. Figure 6 is scaled to show the difference between the values generated from standard case-control process and random sampling regularisation method.

Figure 3 illustrates that the consistency of the top SNPs outlined by standard case-control methods fluctuates across the analyses but present strongly when using the full cohort.

Producing a set of representative features at **stage 4** that are most likely to indicate the presence of a significant relationship is one of the main challenges of this approach. The proposed methodology of this research used both the standard deviation and the mean to produce a unique threshold that takes into consideration the fluctuation of values across random cohorts. Histograms in

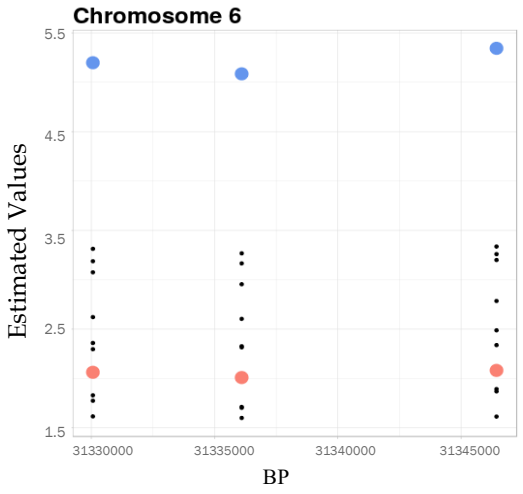

Figure 3: Dot plot comparison of standard case-control vs. random sampling regularisation. The difference between the values produced using standard case-control methods (with genomic control) (blue) , and the values produced by random sampling regularisation analysis, analysis values(black) and mean (pink).

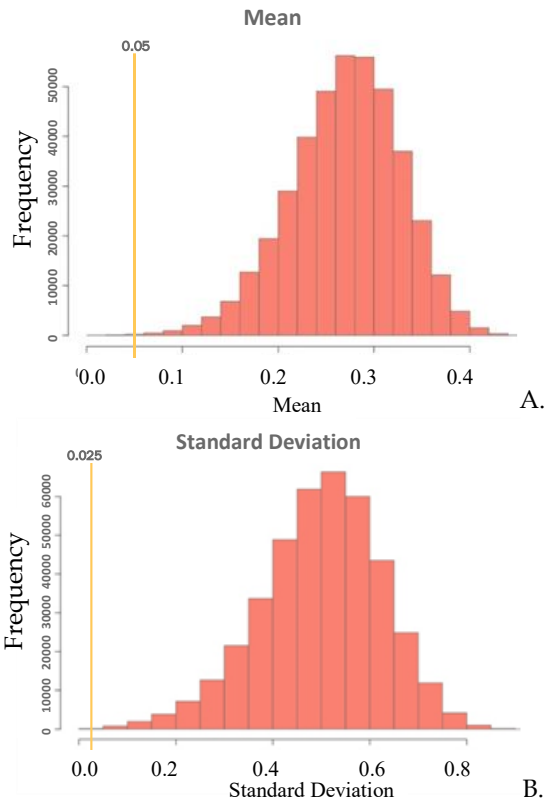

Figure 4: Histograms generated from association analysis using Random Sampling Regularisation showing A. Mean and B. Standard deviation with threshold exclusion measures.

By using the standard deviation values alongside the mean, this approach also considers any SNPs that have an inflated mean due to anomaly results will be excluded based on standard deviation value. The purpose of this feature selection method is to produce a feature set that includes SNPs that show significance but more importantly are consistently significant regardless of the subjects included in the cohort. Stage 4 yielded a total of 57 features for further analysis.

For **stage 5**, features selected using the previous process are inputted into software LAMPlink, using a dominant model. Linkage Disequilibrium pruning is used to remove redundant features that exist as relationship as a result of high LD. A threshold, $\alpha < 0.005$ has been used for the analysis of interaction relationships.

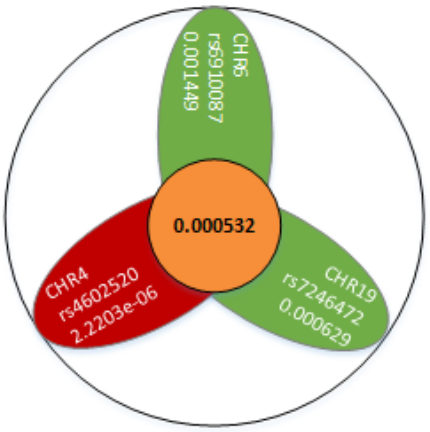

Figure 5: Petal plot for Interaction rs4602520; rs6910087; rs7246472

Adjusted p-values are used to produce petal plots that show the significance value of each singular feature in relation to the overall interaction score. Figure shows the petal plot for interaction rs4602520; rs6910087; rs7246472 where one of the singular SNPs shows a greater p-value than the combined interaction p-value.

In order to focus the results in this section, Table IV and Table V**Error! Reference source not found.** provides the results from the analysis that outline any interaction combinations that produced a significant result of p-value < 0.05 from the training dataset. Having conducted the SNP analysis using the training data, the next process is to analyse the significant combinations outlined using a separate dataset. The purpose of this process is to analyse whether the significant combinations outlined retain significance using an unused set of data which increases confidence in a true positive association. Table VI and Table VII provide the results from the testing set using all significant relationships outlined in Training. Combinations with NA values were omitted as a result of a cell frequency < 10 in a 2x2 contingency table. Combination formatted in bold indicate associations that retained significance using the testing dataset and will therefore for be carried forward.

TABLE IV: STATISTICALLY SIGNIFICANT COMBINATION STATE FROM TRAINING SET DATA

Combination	Variant/s	Results			
		CI <	OR	> CI	OR P
A	1	1.2534	1.3700	1.4975	5.889e-09
	2	1.1518	1.2410	1.3372	1.955e-06
	3	1.1775	1.2811	1.3939	1.371e-06
	4	1.2964	1.5313	1.8088	2.568e-05
		1.0910	1.3880	1.7660	0.02513
		1.0355	1.1246	1.2213	0.01931
		0.7074	0.7561	0.8082	4.984e-12
		1.2726	1.5254	1.8282	0.0001259
		1.1596	1.2858	1.4256	6.228e-05
	5	1.2568	1.5436	1.8958	0.0005136
		1.2544	1.8766	2.8076	0.01016
		1.0808	1.1846	1.2984	0.002373
		0.6836	0.7329	0.7857	1.981e-13
		1.1954	1.4809	1.8346	0.002564
		1.1874	1.3105	1.4464	6.502e-06
	6	1.1737	1.2773	1.3901	1.961e-06
		1.1378	1.9779	3.4383	0.04248
		1.1807	1.3998	1.6595	0.001155
		1.0406	1.1307	1.2286	0.01494
		1.1358	1.7233	2.6149	0.03179
		1.0475	1.1537	1.2706	0.01487
		0.7299	0.7796	0.8326	4.923e-10
	7	2.1493	3.3348	5.1740	6.479e-06
		1.1074	1.7123	2.6478	0.04238
		1.0361	1.1488	1.2737	0.02712
		0.6809	0.7256	0.7733	2.22e-16
		1.9714	3.1958	5.1808	7.632e-05
		1.0411	1.2694	1.5478	0.04778
		1.1656	1.3035	1.4576	9.671e-05
B	8	0.7107	0.7752	0.8456	1.427e-06
	8	0.6919	0.7618	0.8388	3.358e-06
	9	0.6880	0.7579	0.8350	2.492e-06
		0.7009	0.7740	0.8547	2.158e-05
		1.1787	1.2852	1.4014	1.846e-06
C	10	1.1538	1.2961	1.4566	9.028e-06
	11	1.4425	1.8009	2.2484	1.298e-05
		0.6918	0.7442	0.8006	2.949e-11
		1.1505	1.2679	1.3973	5.887e-05
		1.0523	1.1704	1.3017	0.01496
		1.5016	1.9060	2.4194	8.643e-06
D	12	0.7102	0.7743	0.8441	1.11e-06
	13	0.7235	0.7785	0.8377	1.893e-08
	14	0.4953	0.5863	0.6942	1.973e-07
		1.2201	1.3027	1.3908	3.11e-11
		0.4854	0.5807	0.6949	6.284e-07

(Combination A: rs4602520, rs6910087, rs7246472, B: 9q21.13, 9q21.13, C: rs4144827, rs4602520, D: 1q44, rs3924215), Variants 1: rs4602520, 2: rs6910087, 3: rs7246472, 4: rs4602520, rs6910087, 5: rs4602520, rs7246472, 6: rs6910087, rs7246472, 7: rs4602520, rs6910087, rs7246472, 8: 9q21.13, 9: 9q21.13, 9q21.13, 10: rs4144827, 11 : rs4144827, rs4602520, 12: 1q:44, 13: rs3924215, 14: 1q:44, rs3924215)

TABLE V: STATISTICALLY SIGNIFICANT COMBINATION STATE FROM TRAINING SET DATA CONT(1)

Combination	Variant/s	Results			
		CI <	OR	> CI	OR P
A	1	1.1567	1.2465	1.3434	1.269e-06
	2	1.4585	1.7789	2.1697	1.839e-06
	3	2.0276	2.8836	4.1010	7.567e-07
		0.7334	0.7889	0.8487	9.29e-08
		1.8680	2.7163	3.9500	1.135e-05
B	4	1.1994	1.3135	1.4383	7.897e-07
	5	1.2529	1.3695	1.4969	6.17e-09
	6	1.2304	1.3567	1.4959	2.826e-07
		1.2360	1.3677	0.5135	3.638e-07
		0.6890	0.7494	0.8151	1.653e-08
C	7	1.1999	1.3140	1.4390	7.597e-07
	8	2.3032	3.8322	6.3762	1.424e-05
		1.1400	1.2921	1.4644	0.0008
		1.0346	1.1485	1.2749	0.02923
		0.6850	0.7299	0.7777	4.441e-16
		1.0520	1.3280	1.6765	0.04524
D		2.1803	3.4217	5.3699	7.129e-06
		2.2681	3.9384	6.8389	4.392e-05
		1.0970	1.2297	1.3784	0.002906
		1.0311	1.1430	1.2670	0.03282
		0.6985	0.7444	0.7934	2.531e-14
E	9	0.7227	0.7781	0.8377	2.243e-08
	10	0.4105	0.5042	0.6193	2.643e-08
	11	0.2292	0.3538	0.5460	8.201e-05
		0.4477	0.5662	0.7160	6.7e-05
		1.2488	1.3420	1.4422	1.82e-11
F	12	0.4018	0.4952	0.6103	3.178e-08
	13	0.0336	0.1154	0.3965	0.004002
		0.4378	0.5407	0.6678	1.66e-06
		1.6900	1.9632	2.2809	1.397e-13
		0.4266	0.5293	0.6567	1.222e-06
G	14	1.1672	1.2583	1.3564	4.865e-07
	15	1.2947	1.5222	1.7897	1.967e-05
		0.3480	0.5345	0.8209	0.01633

(Combination: A: rs6911024, rs12170250, B: rs6852865, rs4602520, C: rs6852865, rs4602520, rs6910087, rs7246472, D: rs6852865, rs6910087, rs7246472, E: rs3924215, rs6011609, F: 1p12, rs6011609, G: rs4602520, rs6911024, rs7246472), (Variants: 1: rs6911024, 2: rs12170250, 3: rs6911024, rs12170250, 4: rs6852865, 5: rs4602520, 6: rs6852865, rs4602520, 7: rs6852865, 8: rs6852865, rs4602520, rs6910087, rs7246472, 9: rs3924215, 10: rs6011609, 11: rs3924215,rs6011609, 12: 1p12, 13: 1p12, rs6011609,14: rs6911024, 15: rs6911024, rs7246472,16: rs4602520, rs6911024)

TABLE VI: STATISTICALLY SIGNIFICANT COMBINATION STATE FROM TESTING SET DATA

Combination	Variant/s	Results			
		CI <	OR	> CI	OR P
A	1	0.8492	1.0511	1.3016	0.6717
	2	**1.0727**	**1.2425**	**1.4391**	**0.0151**
	3	**1.0538**	**1.2430**	**1.4661**	**0.0302**
	4	0.9018	1.2663	1.7781	0.2526
		0.8138	1.3052	2.0931	0.3537
		1.0148	**1.1918**	**1.3995**	**0.0726**
		0.7422	**0.8464**	**0.9653**	**0.0369**
		0.9308	1.3492	1.9557	0.1845
		0.7689	0.9384	1.1452	0.5995
	5	**0.3166**	**0.4722**	**0.7043**	**0.0020**
		0.5095	1.1699	2.6866	0.7562
		0.9092	1.0881	1.3023	0.4394
		0.8012	0.9191	1.0542	0.3116
		1.2833	**1.9248**	**2.8871**	**0.0079**
		0.6930	0.8424	1.0239	0.1482
	6	**1.0685**	**1.4532**	**1.9763**	**0.0456**
		0.2469	0.6481	1.7015	0.4599
		1.1446	**1.5972**	**2.2288**	**0.0208**
		0.9611	1.1317	1.3327	0.2131
		0.6540	1.6404	4.1149	0.3760
		0.9329	1.1265	1.3603	0.2988
		0.6911	**0.7875**	**0.8972**	**0.0026**
	7	1.1287	2.3549	4.9131	0.0554
		0.4853	1.2741	3.3449	0.6798
		0.8281	1.0138	1.2411	0.9114
		0.7471	**0.8480**	**0.9626**	**0.0323**
		1.2588	**2.9309**	**6.8239**	**0.0364**
		0.7009	1.0675	1.6257	0.7986
		0.6465	0.8036	0.9989	0.0983
B	8	0.7765	0.9204	1.0910	0.4225
	8	0.7921	0.9563	1.1546	0.6967
	9	0.7921	0.9567	1.1556	0.7001
		0.7971	0.9693	1.1787	0.7931
		0.9170	1.0867	1.2878	0.4205
C	**10**	**1.0322**	**1.2517**	**1.5178**	**0.0555**
	11	**1.2900**	**2.0420**	**3.2322**	**0.0106**
		0.7952	0.9189	1.0620	0.3365
		0.7374	0.8914	1.0775	0.3185
		0.8605	1.0655	1.3194	0.6253
		1.2898	**2.1513**	**3.5882**	**0.0138**
D	12	0.7718	0.9137	1.08178	0.3794
	13	0.8863	1.0554	1.2569	0.6114
	14	0.6751	1.0382	1.5967	0.8861
		0.8981	1.0308	1.1830	0.7176
		0.5342	0.8399	1.3206	0.5260

(COMBINATION: A: RS4602520, RS6910087, RS7246472, B: 9Q21.13, 9Q21.13, C: RS4144827, RS4602520, D: 1Q44, RS3924215), (VARIANTS: 1: RS4602520, 2: RS6910087, 3: RS7246472, 4: RS4602520, RS6910087, 5: RS4602520, RS7246472, 6: RS6910087, RS7246472,7:RS4602520, RS6910087, RS7246472, 8: 9Q21.13, 9: 9Q21.13, 9Q21.13, 10: RS4144827, 11: RS4144827, RS4602520, 12: 1Q:44, 13: RS3924215, 14: 1Q:44, RS3924215)

TABLE VII: STATISTICALLY SIGNIFICANT COMBINATION STATE FROM TESTING SET DATA CONT(1)

Combination	Variant/s	Results			
		CI <	OR	> CI	OR P
A	**1**	1.0674	1.2365	1.4325	0.0176
	2	0.7899	0.9144	1.0584	0.3143
	3	0.8285	1.0777	1.4018	0.6397
		0.4149	0.7230	1.2599	0.3368
		NA	NA	NA	NA
B	**4**	**1.1999**	**1.3140**	**1.4390**	**7.597e-07**
	5	0.8365	0.0111	1.2222	0.9239
		NA	NA	NA	NA
		NA	NA	NA	NA
C	6	0.8930	1.0684	1.2783	0.5441
	7	0.8765	1.9787	4.4670	0.1681
		0.6445	0.8196	1.0422	0.1732
		0.7921	0.9567	1.1556	0.7001
		0.7459	**0.8467**	**0.9611**	**0.03069**
		0.4564	0.7541	1.2459	0.3552
D		1.0589	2.2224	4.6647	0.0764
		1.3582	**3.4348**	**8.6863**	**0.0287**
		0.6884	0.8603	1.0752	0.2669
		0.8397	1.0257	1.2528	0.835
		0.7322	**0.8312**	**0.9436**	**0.0165**
E	8	0.8863	1.0554	1.2569	0.6114
	9	0.4947	0.7347	1.0910	0.1996
	10	0.2867	0.6464	1.4575	0.3774
		0.4893	0.7675	1.2037	0.3333
		0.9853	1.1369	1.3117	0.1403
F	**11**	**0.2523**	**0.3939**	**0.6148**	**0.0006**
	12	NA	NA	NA	NA
		0.5325	0.7961	1.1904	0.3512
		1.3164	**1.7737**	**2.3898**	**0.0016**
		0.2694	**0.4224**	**0.6623**	**0.0016**
G	**13**	**1.0684**	**1.2378**	**1.4340**	**0.0171**
	14	1.0473	1.4218	1.9302	0.0583
		0.2432	0.6101	1.5303	0.3768
		0.7358	0.8885	1.0729	0.3024
		1.1105	**1.2653**	**1.4418**	**0.0033**
		0.4618	**0.6428**	**0.8947**	**0.0279**
		0.7494	0.8829	1.0403	0.2116
	15	0.9217	1.2972	1.8257	0.2105
		1.0320	**1.1771**	**1.3425**	**0.0414**
		0.7212	0.8471	0.9950	0.0899
		0.4781	0.7668	1.2297	0.3551
		0.8740	1.0666	1.3017	0.5945
		0.4947	0.7194	1.0461	0.1480

(COMBINATION: A: RS6911024, RS12170250, B: RS6852865, RS4602520, C: RS6852865, RS4602520, RS6910087, RS7246472, D: RS6852865, RS6910087, RS7246472, E: RS3924215, RS6011609, F: 1P12, RS6011609, G: RS4602520, RS6911024, RS7246472), (VARIANTS: 1: RS6911024, 2: RS12170250, 3: RS6911024, RS12170250, 4: RS6852865, 5: RS6852865, RS4602520, 6: RS6852865, 7: RS6852865, RS4602520, RS6910087, RS7246472, 8: RS3924215, 9: RS6011609, 10: RS3924215, RS6011609, 11: 1P12, 12: 1P12, RS6011609, 13: RS6911024, 14: RS6911024, RS7246472, 15: RS4602520, RS6911024)

Any combinations highlighted in grey in Table VI and Table VII remained significant, $p < 0.05$. Further information regarding the real-world statistical significance is shown in Table IX.

Using the combinations outlined from Table VI and Table VII, further analysis is performed to consider their context to real-world information. Using penetrance, incidence and risk to map the extent of their effect, a threshold of penetrance >60% is used to outlined results that present more effective significance to breast cancer. Table IX presents data to outline the penetrance and incidence of each combination. Results outlined in grey surpassed the threshold of 60% penetrance.

TABLE IX: PENETRANCE AND INCIDENCE OF RESULT COMBINATIONS

Interaction	Model	Risk	Pe (%)	I (%)	P
rs6910087	Dominant	1.243	56.4	25.0	0.0151
rs7246472	Dominant	1.243	54.0	18.2	0.0302
rs4602520-rs6910087	AAAG	1.192	56.0	19.4	0.0726
	AAGG	0.846	50.8	63.1	0.0369
rs4602520-rs7246472	Dominant	0.472	51.8	97.0	0.0020
	GAAC	1.925	**67.5**	3.00	0.0079
rs6910087-rs7246472	Dominant	1.453	**61.1**	5.00	0.0456
	AGAC	1.597	**63.3**	4.00	0.0208
	GGCC	0.787	50.0	61.5	0.0026
rs4602520, rs6910087, rs7246472	AAGGCC	0.848	50.4	51.6	0.0323
	GAAGAC	2.931	**76.2**	0.80	0.0364
rs4144827	Dominant	1.252	57.2	12.5	0.0555
rs4144827-rs4602520	Dominant	2.042	**68.9**	2.30	0.0106
	GAGA	2.151	**70.0**	2.00	0.0138
rs6852865-rs4602520-rs6910087-rs7246472	GGAAGGCC	0.847	50.3	50.5	0.0307
rs6852865-rs6910087-rs7246472	AGAGAC	3.435	**79.0**	0.70	0.0287
	GGGGCC	0.831	50.1	52.2	0.0165
1p12	Dominant	0.394	**69.7**	2.50	0.0006
1p12-rs6011609	AAGG	1.774	52.6	95.0	0.0016
	GAGG	0.422	47.0	2.40	0.0016
rs6911024	Dominant	1.237	56.4	25.0	0.0171
rs6911024, rs7246472	AACC	1.265	56.0	38.5	0.0030
	GAAC	0.643	52.0	96.0	0.0279
rs4602520, rs6911024	AAAA	1.177	55.0	37.0	0.0414

Table X presents the statistical characteristics of the top variants and interactions that were identified during this research. Figure 6 provides a visual representation for each interaction based on the OR with upper and lower CI.

TABLE X: PENETRANCE AND INCIDENCE OF TOP RESULT COMBINATIONS

Interaction	State	Risk	RR	Pe (%)	I (%)	P
rs4602520-rs7246472	GAAC	1.925	1.119	67.5	3.00	0.0079
rs610087-rs7246472	Dominant	1.453	1.176	61.1	5.00	0.0456
rs6910087-rs726472	AGAC	1.597	1.219	63.3	4.00	0.0208
rs4602520- rs610087-rs7246472	GAAGAC	2.931	1.460	76.2	0.80	0.0364
rs4144827-rs4602520	Dominant	2.041	1.324	68.9	2.30	0.0106
rs4144827-rs4602520	GAGA	2.151	1.345	70.0	2.00	0.0138
rs6852865-rs6910087-rs7246472	AGAGAC	3.435	1.513	79.0	0.70	0.0287
1p12	Dominant	0.394	0.578	69.7	2.50	0.0006

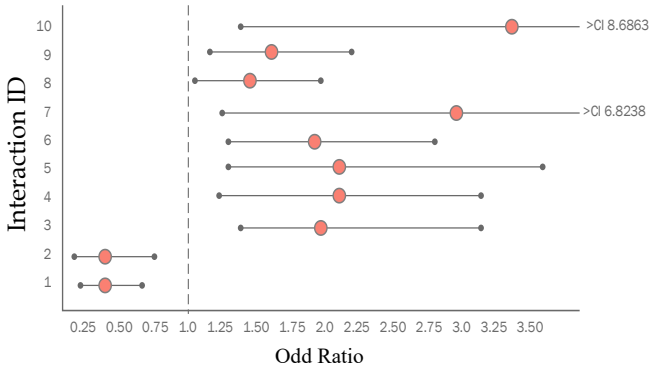

Figure 6: Odds Ratio Tree Plot for Top Results

5. DISCUSSION

The proposed methodology functions as a filter, reducing the feature set through the stages performed. The first stage, QC, used ~500K SNPs and ~28K subjects provided by the DRIVE project. This stage resulted in a dataset of 320,247 features and 13,649 observations. Using the output from QC, training and testing datasets were split 75:25. Random Cohort Sampling was performed to split the training dataset into 9 sizeable subsets of individuals to create a viable averaging sample size for later in further stages. The sample sizes were proportional in the number of cases and controls that were assigned to each subset.

During the QC stage, an association analysis was performed for each of the 9 outlined subsets from previous stages. A further association analysis was performed on the full data output after the QC stage for the purpose of a comparison with standard methods and was further used in the feature selection stage. The results from the association analysis showed a number of suggestive values within the standard GWAS approach; however there were no features that exceeded the genome-wide significance threshold. In Figure 2

Figure a comparison was undertaken to view the difference between the values obtained from the standard GWAS and the proposed 'random sampling regularisation' method. This shows a vast difference between the values obtained by each method that could indicate that either the features from the standard GWAS are inflated or that the values from the 'random sampling regularisation' method are extremely undervaluing the expression of the feature.

During the testing phase, a large majority of the outlined combinations were excluded due to low significance p-value. Penetrance and incidence were computed for the remaining combinations to consider the real-world effect of the combination. Using a lenient threshold of >60%, any combinations that showed a penetrance greater than this threshold were outlined. Of the results, 7 interaction combinations were identified for significance with 1 variant expressing a protective significance.

6. CONCLUSION

A novel methodology was proposed that caters for the needs of epistasis improving flexibility and inspired by random forests machine learning method. The novel methodology outlined in this research presents a statistically conservative approach that

outlines a number of interactions that present viable and reliable options that aim to improve reproducibility by using consistently transparent methods that are fully interpretable. To elaborate, the viability of these variants is conferred by the penetrance and risk, with initial results indicating its relevance using a variety of permutation tests in both training and testing datasets to indicate its significance of <0.05. Reliability is conferred using cascading statistical filters that aim to investigate and reduce the candidate set assuming a null hypothesis. Initial results obtained via Random Sampling Regularisation look promising but should be considered as a preliminary investigation tool. Further validation using another dataset is required to concur reliability of the results. While the performance of the method has been proven significant in this research, there still remain issues that will likely affect outcomes either in a lenient or conservative fashion. One of the most prominent issues is the balance of the standard deviation threshold. While the method is adaptable to specify lenient or conservative thresholds, it is subject to the effects of anomalous data points; this occurrence would present a particular problem in cases were the majority of data points for one variant crowd in a tight cluster with one data point expressing in an anomalous range. The difficulty in addressing this point is the removal of any information could be extracting from the true representation of the variant, or it could be aiding in presenting the true representation of the SNP by removing the data point that has occurred by chance.

Further to this, due to the nature of the method rare alleles may be overlooked during association analysis due to lack of supporting evidence in each subset. Therefore it is proposed that the outlined methodology would perform optimally for complex and common diseases.

ACKNOWLEDGMENT

The dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap.

References

1. S. E. Jackson *et al.*, “Personalised cancer medicine,” *Int. J. Cancer*, vol. 137, no. 2, pp. 262–266, 2015.
2. E. Graham, “Improving Outcomes Through Personalised Medicine,” *NHS Engl.*, pp. 6–10, 2016.
3. Q. Najeeb *et al.*, “Personalized Medicine versus era of " Trial and Error " Abstract : Introduction : Personalized medicine : Pros and Cons,” vol. 19, no. 19, pp. 1–5, 2012.
4. M. E. Lynch *et al.*, “‘One Size Fits All’ Doesn’t Fit When It Comes to Long-Term Opioid Use for People with Chronic Pain,” *Can. J. Pain*, vol. 1, no. 1, pp. 2–7, 2017.
5. Wei-Li Guo, D.S.Huang, “An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency”, Molecular BioSystems, DOI: 10.1039/C7MB00155J, 13(9): 1827-1837, 2017.
6. Zhen Shen, You-Hua Zhang, Kyungsook Han, Asoke K. Nandi, Barry Honig, and D.S.Huang, “miRNA-disease association prediction with collaborative matrix factorization,” Complexity, vol. 2017, no. 2017: 1-9, 2017(SCI).
7. M. Cloitre, “‘one size fits all’’ approach to trauma treatment: Should we be satisfied?,” *Eur. J. Psychotraumatol.*, vol. 6, no. May, 2015.
8. G. Kang *et al.*, “Gene-based Genomewide Association Analysis: A Comparison Study,” *Curr. Genomics*, vol. 14, no. 4, pp. 250–255, 2013.
9. H. Lodish *et al.*, “Protein Sorting Organelle Biogenesis and Protein Secretion,” in *Molecular Cell Biology*, 4th ed., W.H. Freeman and Company, 1999, pp. 675–750.
10. J. Yang *et al.*, “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits,” *Nat. Genet.*, vol. 44, no. 4, pp. 369–375, 2012.
11. Lin Yuan, Chang-An Yuan, and D.S.Huang, “FAACOSE: A fast adaptive ant colony optimization algorithm for detecting SNP epistasis,” Complexity, vol. 2017, no. 2017:1-10, 2017 (SCI).
12. Lin Yuan, Lin Zhu, Wei-Li Guo, Xiaobo Zhou, Youhua Zhang, Zhenhua Huang, and D.S.Huang, “Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping,” IEEE/ACM Transactions on Computational Biology & Bioinformatics, vol. 14, no. 5, pp. 1154-1164. 2017.
13. Su-Ping Deng, Shaolong Cao, D.S.Huang, and Yu-Ping Wang, “Identifying stages of kidney renal cell carcinoma by combining gene expression and DNA methylation data,” IEEE/ACM Transactions on Computational Biology & Bioinformatics, vol. 14, no. 5, pp. 1147-1153. 2017.
14. Lin Zhu Hong, Bo Zhang, D. S. Huang, “Direct AUC optimization of regulatory motifs,” Bioinformatics, 33 (14): i243-i251, 2017, doi: 10.1093/bioinformatics/btx255.
15. Hongbo Zhang, Lin Zhu, D.S.Huang, "WSMD: weakly/supervised motif discovery in transcription factor ChIP-seq data," Scientific Reports, 7, DOI: 10.1038/s41598-017-03554-7, 2017.
16. X. Qu *et al.*, “Association between two CHRNA3 variants and susceptibility of lung cancer: a meta-analysis.,” *Sci. Rep.*, vol. 6, p. 20149, 2016.
17. G. M. Clarke *et al.*, “Basic statistical analysis in genetic case-control studies,” *Nat. Protoc.*, vol. 6, no. 2, pp. 121–133, Feb. 2011.
18. P. M. Visscher *et al.*, “10 Years of GWAS Discovery: Biology, Function, and Translation,” *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, 2017.
19. T. A. Pearson, “How to Interpret a Genome-wide Association Study,” *JAMA*, vol. 299, no. 11, p. 1335, Mar. 2008.
20. Lin Zhu, Su-Ping Deng, D.S.Huang, “Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(2), pp.345-352, 2017.
21. W.-L. Guo, L. Zhu, S.-P. Deng, X.-M. Zhao, and D.S.Huang, “Understanding tissue-specificity with human tissue-specific regulatory networks,” Science China Information Sciences, vol. 59, no. 7, pp. 070105, 2016.
22. Su-Ping Deng, Lin Zhu, and D.S.Huang, “Predicting hub genes associated with cervical cancer through gene co-expression networks,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no.1, pp. 27-35, 2016.
23. Lin Zhu, Weili Guo, Su-Ping Deng, and D.S.Huang, "ChIP-PIT: Enhancing the analysis of ChIP-Seq data using convex-relaxed pair-wise interaction tensor decomposition,"

IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no.1, pp. 55-63, 2016.
24. D.S.Huang, Lei Zhang, Kyungsook Han, Suping Deng, Kai Yang, Hongbo Zhang, " Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," Current Protein & Peptide Science, vol. 15, no. 6: 553-560, 2014.
25. D.S.Huang, Hong-Jie Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.10, no.2, pp.457-467, 2013
26. Chun-Hou Zheng, Lei Zhang, Vincent To-Yee Ng, Simon Chi-Keung Shiu, and D.S.Huang, "Molecular pattern discovery based on penalized matrix decomposition," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.8, no.6, pp.1592-1603, 2011.
27. Chun-Hou Zheng, Lei Zhang, Vincent To-Yee Ng, Simon Chi-Keung Shiu, and D.S.Huang, "Metasample-based sparse representation for tumor classification," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.8, no.5, pp.1273-1282, 2011.
28. C. C. A. Spencer *et al.*, "Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip," *PLoS Genet.*, vol. 5, no. 5, p. e1000477, May 2009.
29. B. Verstockt *et al.*, "Genome-wide association studies in Crohn's disease: Past, present and future," *Clin. Transl. Immunol.*, vol. 7, no. 1, p. e1001, 2018.
30. G. Orozco *et al.*, "Genetics of rheumatoid arthritis: GWAS and beyond," *Open Access Rheumatol. Res. Rev.*, p. 31, Jun. 2011.
31. Lin Zhu, S.P.Deng, D.S.Huang, "A two-stage geometric method for pruning unreliable links in prot ein-protein networks," IEEE Transactions on NanoBioscience, vol. 14, no.5, pp. 528-534, 2015.
32. Su-Ping Deng, Lin Zhu, D.S.Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," BMC Genomics, 2015, 16 (Suppl 3):S4.
33. Su-Ping Deng, D.S.Huang, "SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method," Methods, vol. 69, no. 3: 207-212, 2014.
34. V. Kumar *et al.*, "From genome-wide association studies to disease mechanisms : celiac disease as a model for autoimmune diseases," pp. 567–580, 2012.
35. Zhu-Hong You, Ying-Ke Lei, D.S.Huang, and Xiaobo Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," Bioinformatics, 26(21):2744-2751, 2010.
36. Zhu-Hong You, Zheng Yin, Kyungsook Han, D.S.Huang, and Xiaobo Zhou, "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network", BMC Bioinformatics, doi: 10.1186/1471-2105-11-343, 2010.
37. Jun-Feng Xia, Xing-Ming Zhao, Jiangning Song and D.S.Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," BMC Bioinformatics, vol.11, 174, 2010.
38. Jun-Feng Xia, Xing-Ming Zhao, and D.S.Huang, "Predicting protein-protein interactions from protein sequences using meta predictor," Amino Acids, vol. 39, no.5, pp. 1595–1599, 2010.
39. Chun-Hou Zheng, D.S.Huang, Lei Zhang, and Xiang-Zhen Kong, "Tumor clustering using non-negative matrix factorization with gene selection," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no.4, pp 599-607, 2009.
40. D.S.Huang Xin Huang, "Improved performance in protein secondary structure prediction by combining multiple predictions," Protein and Peptide Letters, vol.13, no.10, pp. 985-991, 2006.
41. D.S.Huang, Xing-Ming Zhao, Guang-Bin Huang, and Yiu-Ming Cheung, "Classifying protein sequences using hydropathy blocks," Pattern Recognition, vol.39, no.12, pp. 2293–2300,2006.
42. D.S.Huang, Chun-Hou Zheng, "Independent component analysis based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol.22, no.15, pp.1855-1862, 2006.
43. P. Kraft *et al.*, "Genetic Risk Prediction — Are We There Yet?," *N. Engl. J. Med.*, vol. 360, no. 17, pp. 1701–1703, Apr. 2009.
44. K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2016," *CA. Cancer J. Clin.*, vol. 66, no. 4, pp. 271–289, 2016.
45. L. Galluzzi *et al.*, "Pathophysiology of Cancer Cell Death," no. January, 2013.
46. D. J. Mcconkeyt *et al.*, "Apoptosis, cancer and' cancer therapy," vol. 6, no. 3, pp. 133–142, 1998.
47. J. Ferlay *et al.*, "Cancer incidence and mortality worldwide : Sources , methods and major patterns in GLOBOCAN 2012."
48. "Breast cancer statistics," *Cancer Research, UK.*, 2018. [Online]. Available: http://www.cancerresearchuk.org/health-professiona. [Accessed: 01-May-2018].
49. "Breast cancer symptoms | Cancer Research UK," *Cancer Research, UK.*, 2018. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/breast-cancer/symptoms. [Accessed: 01-May-2018].
50. C. Harding *et al.*, "Breast cancer screening, incidence, and mortality across US counties," *JAMA Intern. Med.*, vol. 175, no. 9, pp. 1483–1489, 2015.
51. Public Health England, "Clinical guidance for breast cancer screening assessment (NHSBSP Publication No 49)," *Gov.Uk*, no. 4, pp. 1–36, 2016.
52. B. Ardou *et al.*, "Novel Indications for Brca 1 Screening Using Individual Clinical," vol. 267, no. November 1998, pp. 263–267, 1999.
53. E. Gabai-Kapara *et al.*, "Population-based screening for breast and ovarian cancer risk due to *BRCA1* and *BRCA2*," *Proc. Natl. Acad. Sci.*, vol. 111, no. 39, pp. 14205–14210, 2014.
54. S. Shiovitz *et al.*, "Genetics of breast cancer: A topic in evolution," *Ann. Oncol.*, vol. 26, no. 7, pp. 1291–1299, 2015.
55. M.J. Hall *et al.*, "Linkage of early-onset familial breast cancer to chromosome 17q21," Science, vol. 250, pp. 17–22, 1990.
56. D.S.Huang, Systematic Theory of Neural Networks for Pattern Recognition (in Chinese), Publishing House of

Electronic Industry of China, May 1996.
57. D.S.Huang, and Wen Jiang, “A general CPL-AdS methodology for fixing dynamic parameters in dual environments,” IEEE Trans. on Systems, Man and Cybernetics - Part B, vol.42, no.5, pp.1489-1500, 2012.
58. Xiao-Feng Wang, D.S.Huang and Huan Xu, "An efficient local Chan-Vese model for image segmentation," Pattern Recognition, vol. 43, no.3, pp. 603-618, 2010.
59. Bo Li, and D.S.Huang, “Locally linear discriminant embedding: An efficient method for face recognition,” Pattern Recognition, vol.41, no.12, pp. 3813-3821, 2008.
60. Wooster *et al.*, “Localization of a Breast Cancer Susceptibility Gene , BRCA2 , to Chromosome 13q1 2-13,” *Science*, vol. 265, no. September, pp. 2088–2090, 1994.
61. W. Wang *et al.*, “Pathway-based discovery of genetic interactions in breast cancer,” *PLoS Genet.*, vol. 13, no. 9, pp. 1–29, 2017.
62. Q. Milne, R. L., Herranz, J., Michailidou, K., Dennis, J., Tyrer, J. P., Zamora, M. P., ... & Wang, “A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46 450 cases and 42 461 controls from the breast cancer association,” *Hum. Mol. Genet.*, vol. 23, no. 7, pp. 1934–1946, 2013.
63. Y. Sapkota *et al.*, “Assessing SNP-SNP Interactions among DNA Repair, Modification and Metabolism Related Pathway Genes in Breast Cancer Susceptibility,” *PLoS One*, vol. 8, no. 6, pp. 4–9, 2013.
64. J. Namkung *et al.*, “Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method,” *Genet. Epidemiol.*, vol. 33, no. 7, pp. 646–656, Nov. 2009.
65. P. Zeng *et al.*, “Statistical analysis for genome-wide association study,” *J. Biomed. Res.*, vol. 29, no. November 2014, pp. 285–297, Jul. 2015.
66. C. A. Anderson *et al.*, “Data quality control in genetic case-control association studies,” *Nat. Protoc.*, vol. 5, no. 9, pp. 1564–1573, Sep. 2010.
67. A. T. Marees *et al.*, “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis,” *Int. J. Methods Psychiatr. Res.*, vol. 27, no. 2, p. e1608, Jun. 2018.