33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A hybrid framework for Genome Wide Epistasis Discovery

Zehao Tan, Zhuo Zhang, Jiang Liu, Chee Keong Kwoh, Sim Heng Ong, Yik Ying Teo,
Chiea Chuen Khor, E Shyong Tai, Tin Aung, Eranga Vithana and Tien Yin Wong

Abstract—**A hybrid framework integrating Random Forest and Logistic Regression is proposed and implemented for genome-wide epistasis study. The two-stage approach first uses random forest model to capture a pool of epistasis-prone single nucleotide polymorphisms (SNPs), followed by using logistic regression to identify the significant pair-wise epistasis SNPs. We tested the proposed framework on data obtained from Singapore Malay Eye Study (SiMES), in which, 3280 subjects were genotyped on Illumina 610quad arrays and optic nerve parameters were measured in ocular examination. Case-control data set is labeled by choosing the high/low end of vertical Cup-to-Disc ratio (vCDR) values which is a measure of optic nerve degeneration. Our method identified 230 pairs of interacting SNPs with *P-values* below 5×10^{-8}. A preliminary search identified a protein interaction network at a high confidence score of 0.9. The proteins are known to participate in the WNT pathway with involvement in the survival and differentiation of the retina ganglion cells, inferring a strong association with vCDR. The experimental results demonstrate that the proposed framework is valid and efficient for large scale epistatsis study.**

I. Introduction

TRADITIONAL methods of GWAS employ statistical tests to correlate single SNPs to the disease ignoring the possibility of interaction between these SNPs. This is otherwise known as single-locus GWAS. Despite the large number of single locus GWAS carried out, the results of these studies have shown limited success as a tool to identify interactions for complex and rare diseases [1-2]. Given the complicated nature of such diseases, it would be reasonable to postulate that its incidence is correlated to other factors such as gene-gene and gene-environment interactions. This is otherwise known as epistasis. Another reason that prompts the investigation into genetic interaction effects of these diseases is the problem of "missing heritability" which is described by the inter-individual inheritability identified for such diseases being much lower than the actual identified variant for these traits or diseases.

Glaucoma refers to diseases associated to the damage or degeneration of the optic nerve fibers containing the pathways which visual information are be transferred from the eye to the brain. As the second leading cause of blindness worldwide, Open-Angle Glaucoma had been given the nickname "silent thief of sight". Early detection of Open-Angle Glaucoma has been difficult due to its asymptomatic nature. At its early stages open angle glaucoma can be treated with eye drops and medications [3]. GWAS of Glaucoma serves to address the increasing need for improved understanding of the pathophysiology of the disease which is important for its early detection. The fact that one of the established risk factors for Open-Angle Glaucoma is family heritage supports the hypothesis of a genetic correlation to the disease. The complex path-ophysiology of Glaucoma make it reasonable to assume that this genetic correlation is epistatic in nature making it a good candidate of study for our framework.

In our work we choose the definition for epistasis as the statistical deviation from the additive effects of two loci on the phenotype. This is also termed as statistical epistasis [4]. The main challenge in the analysis of epistatic GWAS is the huge amount of computational burden due to the large number of possibilities in terms of the permutation of SNPs. Many statistical and mathematical methods had been introduced to the field of genetics to address the immense computation needs of epistatic GWAS. Numerous reviews for these methods exist [5-7]. Popular methods in these reviews can be categorized by whether it is a parametric method. Parametric method can be subsequently classified as whether it is a partitioning method. Each of these methods has its own limitations and to date there is no single method that is clearly superior to others. Some of these literatures therefore suggest the usage of the hybrid methods capitalizing on the benefits the individual methods to make up for the limitations of another [5]. Through careful survey on the existing knowledge in terms of epistatic GWAS analysis, we propose a framework consisting of two current methods of analysis which are namely Random Forests a recursive partitioning method and Logistic Regression which is a non-parametric technique.

II. Methodology

A. Random Forests

In this method bootstrap sampling is first carried out on the original large dataset and the sample is subjected to splitting into two categories through the choice of a SNP leading to a decision tree [6]. The number of the variables selected from the original dataset per sampling step is known as *mtry*. The split is evaluated using a purity term calculated from the Gini Index at a node *t* which is a measure of how

Manuscript received Apr 15, 2011. This work was supported in part by the Agency for Science, Technology and Research, Singapore, under SERC grant 092-148-0073.

Z. Tan, Z. Zhang and J. Liu are with the Institute for Infocomm Research, A*STAR, Singapore (phone: (065)6408-2163; fax: (065)6776-1378; email: zzhang@i2r.a-star.edu.sg).

C. K. Kwoh is with Nanyang Technology University, Singapore. S. H. Ong, T. Y. Wong, Y. Y. Teo and E. S. Tai are with the National University of Singapore. C. C. Khor is with the Genome Institute of Singapore. A. Tin, E. Vithana and T.Y. Wong are with the Singapore Eye Research Institute.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

well the data is separated in the resulting nodes given by

$$Gini(t) = 1 - \sum_{j=1} p_j^2$$

where p_j is the proportion of category j at node t. The exact criteria of which the quality of the split due to a chosen variable is the associated decrease in impurity, Δi given by:

$$\Delta i = Gini_{parent} - (p_{left} \times Gini_{left} + p_{right} \times Gini_{right})$$

This way each of the SNPs votes for the splitting at each node. These trees are then grown until all the terminal nodes are at its highest purity. After a specified number of trees are generated the algorithm then proceeds to calculate the effect of each of these SNPs in the splitting for the entire "forest" the resulting in the Importance Variable (*VI*) which becomes a measure of the contribution of the particular SNP to the disease in question.

The framework proposed in this paper implements Random Forest through the use of software named Random Jungle (RJ) v1.2.363 by Schwarz *et al* [9]. In terms of *VI* we have chosen to use unscaled permutation importance which calculates and takes the difference between two misclassification rates *(MCR)* for a given predictor variable in a single tree. *MCR1* is calculated from comparing the prediction to the Out-Of-Bag (OOB) samples which are variables that are not chosen in the bootstrap sampling step. *MCR2* is calculated by comparing the prediction to the permutated OOB sample and the recorded permutation importance variables is the averaged increase in the *MCR* due to a given predictor variable. *(ΔMCR=MCR1-MCR2)* [8]. This *VI* is however known to be biased and inflated when the SNPs are in Linkage Disequilibrium (LD). The Random Jungle implementation addresses this problem through the introduction of Conditional Variable Importance *(CVI)* [9]. This was applied by restricting the said permutation to groups of observation which are assigned by analyzing the corresponding dependency structure of the trees grown. As for other parameters, we have chosen number of trees as 10 000 and $mtry = 0.1M$ which is 55783 since M is the total number of predictor variables or SNPs. 30 iterations were conducted to even out variations in *CVI.*

There is however another inherent problem with the use of Random Forests algorithm which remains unsolvable in current implementations. As mentioned a vast amount of data need to be processed in epistatic GWAS analysis, this results in the high dimensionality of the data used. In this case high dimensionality, refers to the $M \approx N$ or $M >> N$ whereby N is the number of samples (individuals). Random Forest predictive ability has been shown to have decreased for such situations.

B. *Logistic Regression*

On the investigation of epistatic GWAS, logistic regression can be modified to include a term to account for the interaction between the SNPs investigated using the following equation:

$$\log\left(\frac{\varphi}{1-\varphi}\right) = \beta_0 + \beta_x X_1 + \beta_y X_2 + \beta_{xy} X_1 X_2$$

Whereby φ represents the probability of the onset of the disease and β_x and β_y are coefficients which indicate the main effects of the SNP 1 and 2 while β_{xy} shows the extent of gene-gene interaction and thus is a measure of epistasis. The logistic regression method detects the association of predictors (SNPs) individually (β_x and β_y) to the outcome (disease) along with the interaction effects (β_{xy}) these predictors have on the outcome. The above equations forms a model are which the data from the large dataset can be tested against it with the null hypothesis being the marker (SNP) chosen has no association to the outcome of interest (disease) and the result is the *P-value* of this test.

As with any single method applied to epistatic GWAS analysis, the main problem with logistic regression is its inability to deal with the high dimensionality of the data. Problems associated to this include higher tendencies to detect false positives and decreased in ability to detect gene-gene interactions [5]. The other problem in the use of logistic regression in detecting epistasis is that high amount of computational time required.

In our proposed framework, logistic regression is carried out using v1.07 of the PLINK software by Purcell *et al* [10]. As with all other implementation of tests for epistasis using logistic regression this software suffers from the large amount time and computational power required for epistatic analysis on whole genome data.

C. *Rationale of hybrid method*

Random Forest on a large sample can be time consuming due to the huge number of trees that needs to be grown and the large number of SNPs drawn in each bootstrap sample for the analysis to be reliable. The combined method reduces the time consumed as we take a significantly smaller number of SNPs from Random Forest to logistic regression. This eliminates the need for results to be conclusive in terms of being accurate to a small number of SNPs. Thus allowing a relaxed requirement in terms of parameters used for RJ. This leads to a reduced the computational time required by the Random Forests segment of the purposed framework.

Our proposed framework then makes use of logistic regression to investigate epistasis as the model for fitting of data which result from Random Forests. The reason for this is that it allows us to avoid the use of Logistic Regression on the sample with large number of SNPs thus significantly reduce the computational time and power required. Since Random Forest tests for association by allowing interaction while Logistic Regression tests for actual interactions, using them in the said sequence allows us to accurately identify the interactions correlated to the disease by looking at a smaller and more important set of data extracted from the original. In addition, our proposed framework also allows us to look at the two locus gene interaction which was not possible through Random Forests. Fig. 1 contains a flowchart of our proposed framework.

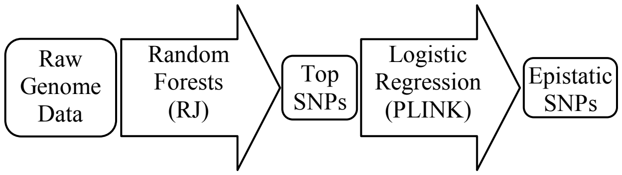

Fig. 1. A flowchart of the proposed framework

D. Data used for our study

To test our proposed framework, we have chosen to use a set of population based GWAS data for Glaucoma named Singapore Malay Eye Study (SiMES). The data is a result of a study was carried out from 2004 to 2007 in an effort to study the causes and risk of blindness and visual impairment in the Singapore Malay community by the Singapore Eye Research Institute (SERI) [11],[12]. The dataset contains autosomal SNP data from 3280 individuals who were genotyped on Illumina 610quad arrays and had optic nerve parameters were measured in ocular examination. Amongst these individuals a set of case and control data is identified through their risk for Glaucoma. This is indicated by the vertical cup-to-disk ratio (vCDR) a parameter measured from retina fundus images. vCDR is a measure of the degeneration of optic nerves and thus represent the degree of risk of an individual to Glaucoma. Individuals with vCDR above 0.65 are considered to have highly degenerated optic nerves thus forming the case in our study while those with vCDR less than 0.3 have optic nerves which are relatively intact forming the control of our study. With this criterion of assessment, a dataset of 233 cases and 458 control subjects with 557,824 SNPs was generated.

III. Results

Since RJ tests for association of the SNPs to the disease allowing interaction, the resulting *CVIs* is an indication of the degree of association to the disease. These *CVIs* can take negative values and zeros which are omitted as this indicates the lack of importance in association to the disease. SNPs with *CVI* as zero may also represent a case which the SNPs has not been selected for any of the trees. An overview of the results of the RJ segment of the framework is shown in the Manhattan Plot of the $-log_{10}(CVI)$ in Fig. 2. SNPs with negative and zero *CVIs* are omitted due to the lack of contribution to the disease investigated. Each one of the 30 iterations of RJ in the hybrid framework took about 17 hours. Tests are performed on a parallel computing cluster with 16 CPUs each with 2 × Quad Core Xeon at 3.0 GHz (X5450) with 8GB of memory running Linux operating system. Much computational time is saved in this aspect as RJ at our parameters would require thousands of iteration for it to be accurate to top few SNPs. For PLINK, our test using $CVI_{10\%}$ took approximately 20 hours while a previous study on 90k SNPs took 14 days [6]. In terms of computational burden, our hybrid method has outperformed the individual methods.

We now need a method to allow us to objectively evaluate the results from RJ to select an appropriate number of SNPs which will be passed on to the next segment of the proposed framework for Logistic Regression carried out by PLINK. For this purpose we propose two methods of selection:

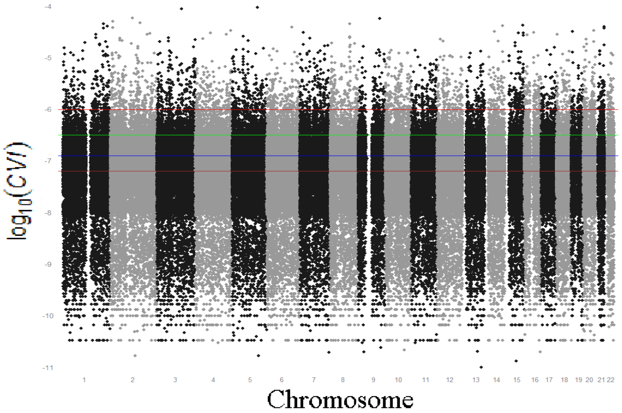

Fig. 2. Manhattan Plot of *CVI* resulting from Random Jungle (30 iterations)

1) Threshold selection through inspection

For this method, we have chosen two values namely $CVI_{Threshold}^{1} = 10^{-6}$ and $CVI_{Threshold}^{2} = 3.16 \times 10^{-6}$. The SNPs above the red lines in Fig. 2 correspond to the SNPs selected by $CVI_{Threshold}^{1}$ (2801 SNPs) and those above the green line were selected by $CVI_{Threshold}^{2}$ (14097 SNPs).

2) Using a percentage threshold from CVI

In this method, we sort the SNPs according their *CVIs* and select the top 10% (55783 SNPs) and 20% (111556 SNPs) which corresponds to the SNPs above the blue and brown line respectively in Fig. 2. The *CVIs* from this criterion of selection shall be referred to as $CVI_{10\%}$ and $CVI_{20\%}$.

Current literature suggests that the minimum *P-value* of 5×10^{-8} is required for a SNP to be considered to have significant correlation to the disease studied for statistical tests employing the Logistic Regression model for GWAS [13]. Using this as a benchmark, we evaluated the results of PLINK in terms of number of significantly interacting SNPs, the results are shown in Table I.

TABLE I
Statistics of Selection Criteria from RJ to PLINK

Selection Criteria	$CVI_{Threshold}^{1}$	$CVI_{Threshold}^{2}$	$CVI_{10\%}$	$CVI_{20\%}$
No. of significantly interacting SNP pairs	0	10	72	158
No. of SNP pairs on coding DNA	0	0	12	27
Lowest *P-value*	10^{-7}	10^{-9}	10^{-10}	10^{-11}

IV. Discussion

To validate the results of our proposed framework, we looked up an online database named dbSNP which allow us to map the SNPs to the genes. Logically a SNP needs to be on a gene which codes for protein for it exhibit epistasis. Therefore pairs with either SNP on noncoding regions of the DNA are considered to be false positives, shown in Table II.

If the proteins coded by these genes interact, that would give strong evidence supporting the discovery of the detected pairs of epistasis SNPs. To do so, we first perform a search using the gene symbols resulting from dbSNP on an established protein-protein interaction database. For the purpose of our work, we have chosen to use STRING [14]. Amongst the 12 pairs of SNPs identified using the $CVI_{10\%}$, 1 pair of interacting SNPs is found to be correlated to a protein-protein interaction network with a highest confidence score of 0.9. Within the 27 pairs of SNPs identified using the $CVI_{20\%}$, 2 pairs of interacting SNPs are found to have interaction with a lower confidence score of 0.7. The statistics of these interacting pairs are shown in Table II. This suggests that $CVI_{10\%}$ may be superior in identifying epistasis with stronger evidence in terms of protein-protein interaction while increasing the number of SNPs (i.e. from $CVI_{10\%}$ to $CVI_{20\%}$) increases the absolute number of interactions detected. Comparing the interaction *P-values* to single locus *P-values* in Table II we can also see that our hybrid method allows the identification of SNPs which are not identified by single locus logistic model.

TABLE II
STATISTICS OF INTERACTING SNPS

	SNP 1	SNP 2	**Interaction *P-value***	CVI
Chromosome	5	13		10%
SNP id	rs152402	rs797208		
Gene Symbol	TCF7	STARD13		
P-value	**1.633×10^{-01}**	**4.049×10^{-02}**	**2.458×10^{-08}**	
Chromosome	3	5		20%
SNP id	rs9843488	rs10062069		
Gene Symbol	PDCD6IP	FER		
P-value	**1.731×10^{-02}**	**1.011×10^{-02}**	**3.320×10^{-08}**	
Chromosome	3	13		20%
SNP id	rs2970535	rs7331661		
Gene Symbol	CACNA2D3	TBC1D4		
P-value	**6.864×10^{-01}**	**4.904×10^{-01}**	**4.903×10^{-08}**	

The improvement of the hybrid method over the individual methods of GWA can be shown by the results of comparison in Table III. The time taken for analysis was used as a measure of the improvement of computational efficiency. Our method has achieved the lowest time taken for epistatic analysis. The time taken for RF using RJ is parameter dependent, an accurate analysis on our dataset is estimated to take several weeks. In terms of ability to detect epistasis we compare the *P-values* and protein interaction network detected. The hybrid method achieved the best performance with a *P-value* lower than single locus LR and the most number of protein interaction network detected.

To prove the correlation of the protein interaction network with Glaucoma we refer to a genetic study done by Wang *et al*. in 2008, in which elevated expression of WNT antagonist is correlated to the increased IOP leading to Glaucoma. Several of the genes in protein interaction network identified are involved in the WNT expression pathway [15]. These are β-catenin (CTNNB1) which is a key WNT intermediate signaling molecule and transcription factor (TCF7) that mediate WNT-regulated gene expression which plays a large role in the survival and differentiation of retina ganglion cells [16]. This gives the validation on the ability of the hybrid framework in detecting epistasis in large scale studies, as it is able to identify part of a protein interaction network which is physiologically proven to be related to the maintenance of the optic nerve cells and thus Glaucoma.

TABLE III
COMPARISON OF METHODS OF ANALYSIS

Method	*Single Locus (LR)*	*Epistatic (LR)*	*RF*	*Hybrid*
Computation Time	< 5 minutes	>> 14 days	#	< 7days
Lowest *P-value*	10^{-7}	N/A	-	10^{-11}
Number of protein interaction networks	N/A	N/A	1*	2*/1

*identified at low (< 0.9) confidence score, #Parameter dependent

REFERENCES

[1] C. S. Ku, E. Y. Loy, Y. Pawitan and K. S. Chia, "The pursuit of genome-wide association studies: where are we now?," J Hum Genet, vol. 55, no. 4, pp. 195-206, 2010.

[2] J. Hardy and A. Singleton, "Genomewide association studies and human disease," N Engl J Med, vol. 360, no. 17, pp. 1759-1768, 2009.

[3] R. P. Crick, "Early detection of glaucoma," Br Med J, vol. 285, no. 6348, pp. 1063-1064, 1982.

[4] R. Fisher and others, "The correlation between relatives on the supposition of Mendelian inheritance," Transactions of the Royal Society of Edinburgh, vol. 52, pp. 399-433, 1918.

[5] A. G. Heidema, J. M. A. Boer, N. Nagelkerke, E. C. M. Mariman, D. L. van der A and E. J. M. Feskens, "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases," BMC Genet, vol. 7, no. 1, pp. 23, 2006.

[6] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," Nat Rev Genet, vol. 10, no. 6, pp. 392-404, 2009.

[7] R. M. Cantor, K. Lange and J. S. Sinsheimer, "Prioritizing GWAS results: A review of statistical methods and recommendations for their application," Am J Hum Genet, vol. 86, no. 1, pp. 6-22, 2010.

[8] Y. Sun, "Multigenic modeling of complex disease by random forests," Adv Genet, vol. 72, pp. 73, 2010.

[9] D. F. Schwarz, I. R. König 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, 2010.

[10] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira and D. Bender *et al*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," Am J Hum Genet, vol. 81, no. 3, pp. 559-575, 2007.

[11] A. W. P. Foong, S. M. Saw, J. L. Loo, S. Shen, S. C. Loon and M. Rosman *et al*, "Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES)," Ophthalmic Epidemiol, vol. 14, no. 1, pp. 25-35, 2007.

[12] Z. Zhang, J. Liu, C. Kwoh, X. Sim, W. Tay and Y. Tan *et al*, "Learning in Glaucoma Genetic Risk Assessment," in Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, pp. 6182-6185, 2010.

[13] I. Pe'er, R. Yelensky, D. Altshuler and M. J. Daly, "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants," Genet Epidemiol, vol. 32, no. 4, pp. 381-385, 2008.

[14] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth and P. Minguez *et al*, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Res, vol. 39, no. Database issue, pp. D561-D568, 2011.

[15] W. H. Wang, L. G. McNatt, I. H. Pang, J. C. Millar, P. E. Hellberg and M. H. Hellberg *et al*, "Increased expression of the WNT antagonist sFRP-1 in glaucoma elevates intraocular pressure," J Clin Invest, vol. 118, no. 3, pp. 1056-1064, 2008

[16] M. A. Fragoso, H. Yi, R. E. I. Nakamura and A. S. Hackam, "The Wnt Signaling Pathway Protects Retinal Ganglion Cell 5 (RGC-5) Cells from Elevated Pressure," Cell Mol Neurobiol, vol. 31, no. 1, pp. 163-173, 2011.