I J B R
International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 2, 2011, pp-227-235
Available online at http://www.bioinfo.in/contents.php?id=21

THE USE OF THE ANT COLONY ALGORITHM FOR THE DETECTION OF MARKER ASSOCIATIONS IN THE PRESENCE OF GENE INTERACTIONS

ROBBINS K.[1], BERTRAND K.[1] AND REKAYA R.[1,2,3*]
[1] Department of Animal and Dairy Science, University of Georgia, Athens, USA
[2] Department of Statistics, University of Georgia, Athens, USA
[3] Institute of Bioinformatics, University of Georgia, Athens, USA
* Corresponding Author: Email: rrekaya@uga.edu

Received: September 29, 2011; Accepted: October 29, 2011

Abstract - In recent years there has been much focus on the use of single nucleotide polymorphism (SNP) fine genome mapping to identify causative mutations for traits of interest; however, many studies focus only on the marginal effects of markers, ignoring potential gene interactions. Simulation studies have shown that this approach may not be powerful enough to detect important loci when gene interactions are present. Although several attempts have been made to study potential gene interaction, the number of SNP markers considered in these studies is often limited. Given the prohibitive computation cost of modeling interactions in studies involving a large number SNP, there is a need for methods that can account for potential gene interactions in a computationally efficient manner to be developed. In this study, the ant colony optimization algorithm (ACA) and logistic regression on large number of SNP genotypes were used. Our procedure was compared to sliding window (SW/H), and single locus genotype association (RG) methods used in haplotype analyses. A binary trait simulated using an epistatic model and HapMap ENCODE SNP genotypes was used to evaluate each algorithm. Results show that the ACA outperformed SW/H and RG under several simulation scenarios, yielding substantial increases in power to detect genomic regions associated with the simulated trait.
Key Words: Ant Colony, Marker association, Gene interaction

Introduction

With the advent of high-throughput, cost effective genotyping platforms, there has been much focus on the use of high-density single nucleotide polymorphism (SNP) genotyping to identify causative mutations for traits of interest, and while putative mutations have been identified for several traits, these studies tend to focus on SNP with large marginal effects [1, 2]. However, several studies have found that gene interactions may play important roles in many complex traits [3, 4]. Given the high density of SNP maker maps, examining all possible interactions is seldom possible computationally. As a result, studies examining gene interactions tend to focus on a small number of SNP, previously identified as having strong marginal associations.

While this approach has shown some success, simulation studies conducted by [5] and [6] showed that, in the presence of several types of gene interactions, there is reduced power to detect causative loci with models estimating only marginal effects. Using an exhaustive search of all two-way interactions, Marchini et al. achieved greater power to detect causative mutations than when estimating only marginal effects. Due to the high computational cost of this approach, a two-stage model was proposed, in which SNP were selected in the first stage based on marginal effects an then tested for interactions in the subsequent stage [5]. This approach could, however, result in the failure to detect important regions of the genome in the first stage of the model. As such, there is a need for methodologies capable of identifying important genomic regions in the presence of potential gene interactions when large numbers of markers are genotyped.

Given that the examination of all possible SNP interactions is computationally infeasible with dense SNP marker maps covering large regions of a genome, an alternative approach must be considered. One approach would be to view the identification of groups of interacting SNP as an optimization problem, for which several algorithms have been developed. These algorithms are designed to search large sample spaces for globally optimal solutions and have been applied to a wide range of problems [7, 9]. Through the evaluation of groups of loci efficiently selected from different regions of the genome, optimization algorithms should be able to account for potential interactions. Kooperberg et al. [10] utilized an optimization algorithm, referred to as simulated annealing (SA), to examine interaction effects; however, only 32 SNP were considered in the model selection process. For studies involving hundreds or even thousands of SNP, efficient algorithms are needed to search the sample space for optimal solutions.

One algorithm, the ant colony algorithm (ACA), has been shown to be efficient in high-dimension data sets [11]. The ACA, developed by Dorigio and Gambardella [12], is based on the mechanism by which ant colonies find the shortest route to a food source. Ants communicate through a chemical pheromone trail, deposited as they transverse a given path. Ants that choose a shorter path will transverse the distance at a faster rate, thus depositing more pheromone in the process. As the pheromone builds, ants will begin to preferentially choose the shorter path leading to a positive feed back system. Dorigio and Gambardella [12] showed that the communication between ants had a synergistic effect allowing the ACA to reach optimal solutions in fewer iterations when compared to other optimization algorithms. In the case of SNP association studies, the 'path' is represented by a selected subset of SNP markers, and performance is evaluated based on the fit of a logistic regression for binary traits.

For this study, a modified ACA, enabling the use of permutation testing for global significance, was combined with logistic regression and implemented on a simulated binary trait under the influence of interacting genes. The performance of the ACA was evaluated and compared to models accounting for only marginal effects.

Materials and Methods

Logistic regression: Groups of SNP markers were evaluated based in haplotype genotype effects estimated as log odds ratios (*lor*) using logistic regression (LR). The relationship between the *lor* and the binary response can be expressed as:

$$y_i = \begin{cases} 1 & \text{if } lor_i \geq 0 \\ 0 & \text{if } lor_i < 0 \end{cases}$$

The log odds ratio lor_i is modeled as:

$$lor_i = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\boldsymbol{\beta} + e_i \qquad (1)$$

where P_i = probability ($y_i = 1$) and **X** is a matrix containing indicator variables for the haplotypes formed from the selected SNP. Groups of SNP markers with less than two corresponding observations were discarded, and analysis was conducting on all remaining marker groups. The link function of the log odds ratio $\mathbf{X}_i\boldsymbol{\beta}$ with the binary response y_i gives the following equations**:**

$$p_i(y_i = 0) = \frac{1}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \quad \text{and}$$

$$p_i(y_i = 1) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \qquad (2)$$

yielding the following relationships:

$$y_i = \begin{cases} 1 \text{ if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} \geq 0.5 \\ 0 \text{ if } \dfrac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})} < 0.5 \end{cases}$$

Marginal effects model: The genotype and haplotype association methods were implemented using R functions developed by [13, 14]. The haplotype analysis was implemented using a sliding window approach which utilizes a window of *k* SNP in width sliding across the genome *h* SNP at a time. Individual SNP scores were determined as the maximum average of all haplotypes containing a given SNP.

Ant colony algorithm: The ACA employs artificial ants that communicate through a probability density function (PDF) that is updated at each iteration with weights or "pheromone levels", which are analogous to the chemical pheromones used by real ants. In the case of SNP association studies, the weights can be determined by the strength of the association between selected haplotypes or genotypes and the trait of interest. Using the notation in [12, 15], the probability of sampling SNP *m* at time *t* is defined as:

$$P_m(t) = \frac{(\tau_m(t))^{\alpha}\eta_m^{\beta}}{\sum_{m=1}^{nf}(\tau_m(t))^{\alpha}\eta_m^{\beta}} \qquad (3)$$

where $\tau_m(t)$ is the amount of pheromone for SNP *m* at time *t*; η_m is some form of prior information on the expected performance of SNP *m*; α and β are parameters determining the weight given to pheromone deposited by ants and a priori information on the features, respectively

Using the PDF as defined in equation (4), each of *j* artificial ants will select a subset S_k of *n* SNP from the sample space S containing all SNP. Given the relationship between adjacent SNP, ants can randomly change SNP selections following a multinomial distribution. Changes in SNP selection are limited to the three adjacent SNP on either side of the originally selected SNP marker. The pheromone level of each feature *m* in S_k is then updated according to the performance of S_k as:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \qquad (4)$$

where ρ is a constant between 0 and 1 representing the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$ is the change in pheromone level for feature *m* based on the sum of accuracy of all S_k containing SNP *m*, and is set to zero if SNP *m* was not selected by any of the artificial ants.

While the algorithm, in the aforementioned form, can be used to subjectively identify markers, it is not well suited for the calculation of permutation p-values. When updating the pheromone function, as previously described in equation (4), the final pheromone levels are relative not only to prediction accuracy, but the number of times a SNP marker is selected. As a result, the amount of pheromone deposited on a feature depends greatly the amount of pheromone deposited on all other SNP markers and can vary wildly from permutation to permutation. One obvious solution to this problem is to

use the average accuracy of all S_k containing genotypes for SNP *m*; however, this approach substantially reduces the ACA's ability to efficiently burn in on good solutions, an attribute needed to detect unknown gene interactions in high-dimension data sets.

To overcome these limitations, a two-layer pheromone function was developed:

$$P_m(t)=\frac{\tau_m(t)^{\alpha}\tau 2_m(t)^{\alpha 2}\eta_m{}^{\beta}}{\sum_{m=1}^{nf}\tau_m(t)^{\alpha}\tau 2_m(t)^{\alpha 2}\eta_m{}^{\beta}} \quad (5)$$

where $\tau_m(t)$ is the first pheromone layer updated using the sum of accuracies for all S_k containing SNP *m*; $\tau 2_m(t)$ is the second pheromone layer updated using the average accuracy of all S_k containing genotypes for SNP *m*; and η_m, α, β are as previously described. For the current study, α and $\alpha 2$ were set to 1, β was set to .3 and the prior information (η_m) was the prediction the accuracy of SNP marker *m*, obtained using logistic regression on genotypes.

The pheromone for $\tau_m(t)$ was updated using equation (4) and $\tau 2_m(t)$ was updated using the following equation:

$$\tau 2_m(t+1)=[t*\tau_m 2(t)+\Delta\tau_m 2(t)]/(t+ns) \quad (6)$$

where *t* is the iteration number; $\Delta\tau_m 2(t)$ is the change in pheromone level for feature *m* based on the sum of accuracy of all S_k containing genotypes for SNP *m*, and is set to zero if feature *m* was not selected by any of the artificial ants; and ns is the number of times SNP *m* was selected at iteration *t*. Permutation p-values were calculated using $\tau 2_m(t)$ only.

The procedure can be summarized in the following steps:

1. Each ant selects a predetermined number of SNP markers.
2. Using the selected SNP markers, accuracies are computed using logistic regression on haplotypes or genotypes.
3. The pheromone for each selected group of SNP, S_k, is calculated as:

$$pheromone_k = acc^{(1-acc)} \quad (6)$$

4. The change in pheromone at time *t* is then calculated using equations (4) and (6).
5. Following the update of pheromone levels according to equations (4) and (6), the PDF is updated according to equation (5) and the process is repeated until pheromone levels have converged.

Data simulations: Genotype data on 90 unrelated individuals from the Japanese and Han Chinese populations were downloaded from the HapMap ECODE project website. Each simulation scenario was replicated five times using two 500 Kbp regions on chromosome 2, comprising 2047 polymorphic SNP. All SNP haplotypes were assumed to be known with out error. The binary disease trait was simulated under a two locus epistatic model as seen in Table 1. The loci of the causative mutations were selected at random; with the frequencies of the causative mutations being .58 and .6. Although these frequencies might be considered high, it was necessary to restrict selection to SNP with mutant allele frequencies greater than .5. This was done to insure a reasonable simulated disease incidence of 15%. A plot illustrating the LD of all SNP with the two causative mutations is shown in Fig (**1**). The plot shows a large peak of high LD with rs2049736 (SNP 409), while the peak of high LD with rs28953468 (SNP 2041) is substantially narrower, and is preceded by a plateau of SNP in moderate LD with rs28953468.

Permutation testing was used to access global significance for all models used in the study. Statuses were randomly shuffled amongst subjects, with haplotype effects, genotype effects and association p-values re-estimated for each new configuration of the response variables. The largest estimated haplotype/genotype effect or the smallest haplotype/genotype association p-value from each permutation was saved to form an empirical distribution used for calculation of p-values. One hundred permutations were performed, yielding p-values accurate to 1%. Power was calculated as the proportion of times a given method identified at least one SNP marker in high LD ($r^2 \geq .80$) with a causative mutation.

Results

Estimates of power for the three methods can be found in Table 2. Methods employing the ACA showed substantial increases in power when compared to the methods accounting for only marginal effects. Due to the fact that the trait was simulated under a dominance model, analysis of genotypes tended yielded superior results when compared to haplotype analysis. Despite the inherent advantage of genotype analysis using a dominance model, the ACA using haplotypes (ACA/H) still showed greater power than RG/D in both scenarios. For scenario 2, all models showed a reduction in power; however, the superiority of the ACA methodologies remained constant, with the ACA using LG on genotypes assuming a dominance model (ACA/G/D) yielding 66.7% increase in power for both scenarios when compared to the next best method, RG/D.

Plots of the associative effects, obtained using SW/H, ACA/G/D, and RG/D, are shown in Fig. (**2**) and (**3**). When compared to the LD plot (Fig. (**1**)) all methods show good correspondence for scenario 1, though only the ACA/G/D was able to identify markers for both causative mutations in all replicates. In scenario 2, where the genetic effect was greatly reduced, plots of associative effects tended to be noisier for all models, with the ACA/G/D again showing superior performance, identifying several SNP markers having only moderate LD with causative mutation rs28953468.

To determine the effectiveness of the permutation on pheromone levels, the cumulative distribution, based on LD with causative mutations, of SNP identified as being

significantly associated with simulated trait by ACA/G/D and RG/D were plotted and can be found in Fig. (**4**). Despite similarities in the average number of SNP identified by ACA/G/D (15.4) and RG/D (22), the distributions of these SNP, differed substantially. In contrast to RG/D, the ACA/G/D identified a large number of SNP having LD between .35-.45. These SNP corresponded to the broad plateau of SNP in LD with SNP 2041. Unlike RG/D, the ACA/G/D also identified several SNP (5.19%) having less than .10 LD with either of the causative mutations, an unexpected result given the strict family-wise significance thresholds (α=0.05) imposed on all models. Surprisingly, both methodologies identified a large number of SNP having LD of approximately ~.2. Upon closer examination it was found that these SNP had LD of ~.2 with both causative mutations, likely artifacts of the data resulting from the relatively small sample size. The LD with both causative mutations imparted a portion of the epistatic effect on these SNP, resulting in significant associations with the simulated traits.

Discussion

The substantial increase in power observed when using ACA/LR demonstrates the effectiveness of the ACA in accounting for epistasis. The first layer of the pheromone function allows the ants to burn-in on optimal groups of SNP, in this case, SNP epistatically interacting. The second layer of the pheromone function yields a measurement of accuracy for a given SNP, accounting for its interaction with other SNP loci. Initially these loci are randomly selected based only on marginal effects; however, as the algorithm burns in, the interacting SNP begins to selected together more frequently, increasing the contribution of the epistatic effects to the pheromone used for permutation testing. For the scenarios simulated in this study, this positive feedback allowed the ACA to efficiently identify SNP in high to moderate LD with the causative mutations that had no significant marginal effects, as evidenced by the decreased power of the marginal effects models.

The relatively high error rates observed when using the ACA/G/D were somewhat surprising given the strict control placed on family-wise error. One would expect, given the number of replications, that detection of false positives would be between 0 and 1, in this study. The ACA/G/D detected significant associations for a total of 8 SNP markers having LD less than .1 with causative mutations. One possible explanation could be the small number of permutations conducted, as 100 permutations yield p-values accurate to only one tenth. The use of 500 permutations would be more adequate; unfortunately, given the high number of replicates conducted in this study, 500 permutations would be too computationally costly.

Regardless of the cause, a false positive rate of 5.19% would generally be considered acceptable in high-dimension association studies for which the goal was to identify a small subset of markers for further evaluation [16], especially when considering the increase in power obtained when using the ACA/LR. However, given the larger number of SNP in low LD with causative mutations identified by ACA/G/D when compared to other methods, there was concern that the increased power associated with the ACA could be the result of less stringent thresholds being applied to the ACA. To belay these concerns a more stringent threshold (α=0.03) was applied to the ACA/G/D. Using this threshold, only one SNP, having LD less than 0.1 with the causative mutations, was identified. While power was slightly reduced, ACA/G/D still yielded increases in power of 50.0% and 33.3% over RG/D for scenarios 1 and 2, respectively.

Since association studies involving large numbers of SNP are generally exploratory in nature, the number of potentially interacting SNP would be unknown. This would necessitate the ACA be robust relative to the number of SNP used to form groups and the number of SNP interacting to control the trait of interest. In this regard the number of SNP used by the ACA, relative to the number of SNP interacting in the simulated model, did show variation, but this variation showed no discernable trends. Although these results suggest some level of robustness, several runs of the algorithm, selecting various group sizes of SNP formation, might best insure that optimal results are reached, a practice often used for analysis of haplotypes using sliding windows.

Conclusion

In the presence of simulated epistasis, the proposed optimization methodology obtained substantial increases in power, demonstrating the effectiveness of machine learning approaches for the analysis of marker association studies in which gene interactions may be present. Although the ACA methods identified more SNP markers that could be construed as false positives, the use of a more stringent threshold eliminated the problem without greatly reducing the advantage of the ACA, in terms of power, when compared to other methods. The results of this study provide compelling evidence that methodologies capable of efficiently modeling gene interactions, such as the model proposed in this study, could yield superior performance detecting important SNP markers for complex traits.

References

[1] Hugot J.P., Chamaillard M., Zouali H., Lesage S., Cézard J.P., Belaiche J., Almer S., Tysk C., O'Morain C.A., Gassull M., Binder V., Finkel Y., Cortot A., Modigliani R., Laurent-Puig P., Gower-Rousseau C., Macry J., Colombel J.F., Sahbatou M., Thomas G. (2001) *Nature*. 411(6837), 599-603.

[2] Woon P.Y., Kaisaki P.J., Braganca J., Bihoreau M.T., Levy J.C., Farrall M. and Gauguir D. (2007) *Proc. Natl. Acad. Sci.*, 104(36), 14412-14417.

[3] Coutinho A.M., Sousa I., Martins M., Correia C., Morgadinho T., Bento C., Marques C.,

Ataíde A., Miguel T.S., Moore J.H., Oliveira G., Vicente A.M. (2007) *Hum. Genet*., 121(6), 243-256.

[4] Barendse W., Harrison B.E., Hawken R.J., Ferguson D.M., Thompson J.M.,Thomas M.B. and Bunch R.J. (2007) *Genetics,* 176(4), 2601-2610.

[5] Marchini J., Donnelly P., Cardon L.R. (2005) *Nat. Genetics*, 37(5), 413-417.

[6] Pickrell J., Clerget-Darpoux F., Bourgain C. (2007) *Genet. Epidemiol*. 31(7), 748-762.

[7] Shymygelska A. and Hoos H.H. (2005) *BMC Bioinformatics*. 6(30), 1-22.

[8] Kreiger M.J.B., Billeter J.B. and Keller L. (2000) *Nature.* 406(6799), 992-995.

[9] Ding Y.P., Wu Q.S. and Su Q.D. (2005) *Analytical Sciences.* 21(1), 327-330.

[10] Kooperburge C., Bis J.C., Marciante K.D., Heckbert S.R., Lumley T. and Psaty B.M. (2006) *Am. J. Epidemiol*., 165(3), 334-343.

[11] Robbins K.R., Zhang W., Rekaya R., and Bertrand J.K.. (2007) *Math Med Biol*., 24(4), 413-426.

[12] Dorigio M. and Gambardella L.M. (1997) *BioSystems*. 43(2), 73-81.

[13] Gonzalez J.R., Armengol L., Sole X., Guino E., Mercader, J.M., Estivill X. and Moreno V. (2007) *Bioinformatics*. 23(5), 644-645

[14] Sinnwell J.P. and Schaid D.J. (2005) *R package version* 1.2.2.

[15] Ressom H.W., Varghese R.S., Orvisky E., Drake S.K., Hortin G.L., Abdel-Hamid M., Loffredo C.A. and Goldman R. (2007) *Bioinformatics*, 23(5), 619-626.

[16] Benjamini Y. and Yekeli D. (2005) *Genetics.* 171(2), 783-790.

Table-1 - Relative risk for simulated trait[a].

	Scenario 1				Scenario 2			
	AB	aB	Ab	ab	AB	aB	Ab	ab
AB	1	1	1	1	1	1	1	1
aB	1	1	1	1	1	1	1	1
Ab	1	1	1	1	1	1	1	1
Ab	1	1	1	15	1	1	1	10

[a] Risks are relative to the aa/bb genotype.

Table 2 - Power calculations[a].

	Scenario 1			Scenario 2		
	1 locus	2 locus	3 locus	1 locus	2 locus	3 locus
ACA/G/D	___	1.00	0.90	___	0.50	0.40
ACA/G/C	___	0.70	0.80	___	0.40	0.40
ACA/HAP	___	0.60	0.70	___	0.50	0.40
RG/D	0.60	___	___	0.30	___	___
RG/C	0.30	___	___	0.30	___	___
SW/HAP	___	0.10	0.20	___	0.00	0.00

[a] Power was calculated as the proportion of times at least one SNP in high linkage disequilibrium (>.8) with a causative mutations was detected by the model at α=.05 for genome-wide significance.

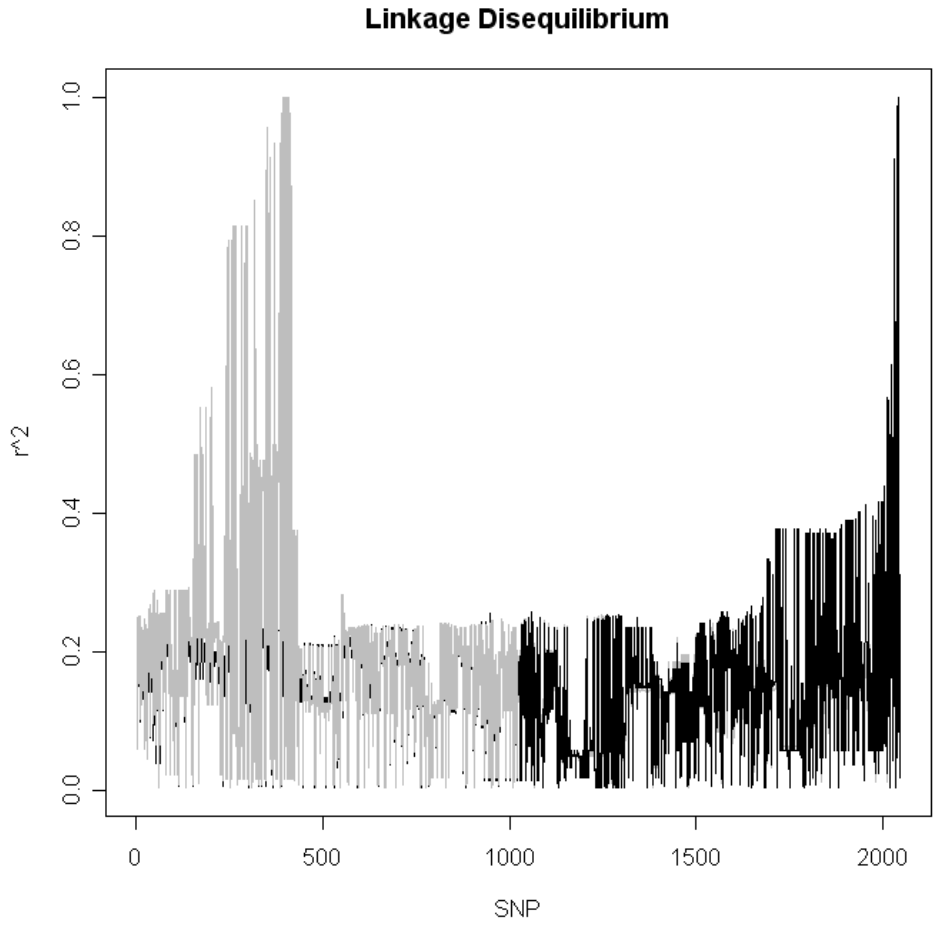
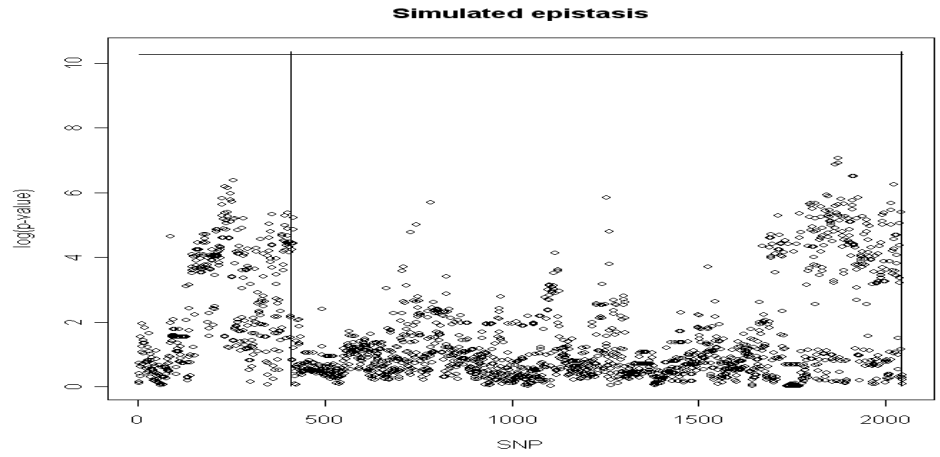
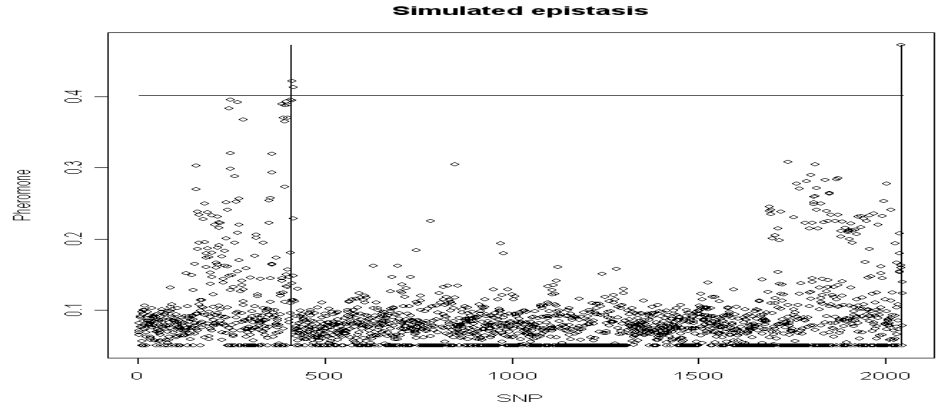


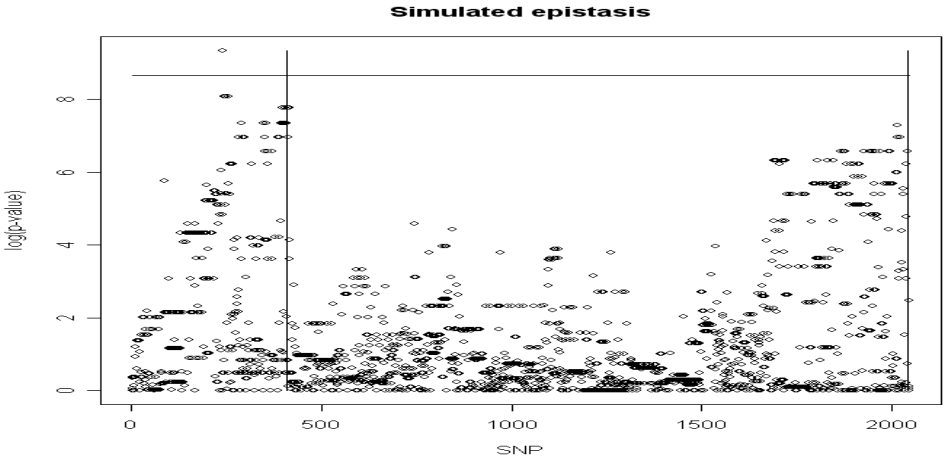
Fig.1-Plots of each marker's linkage disequilibrium (LD) with the two causative mutations. The light grey line represents LD with the causative mutation located at position 409. The black line represents LD with the causative mutation located at position 2041.



2 (a)

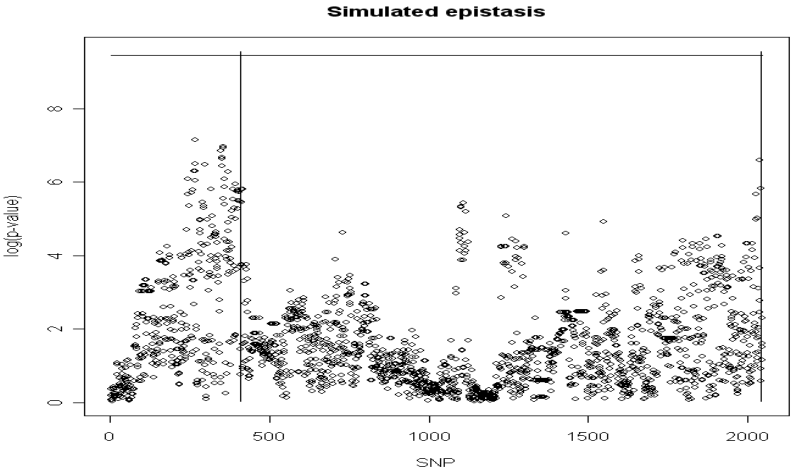


2 (b)

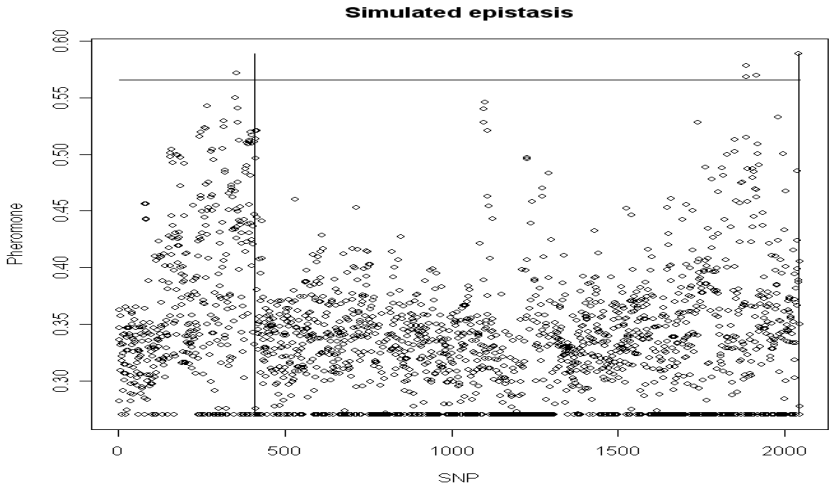


2 (c)

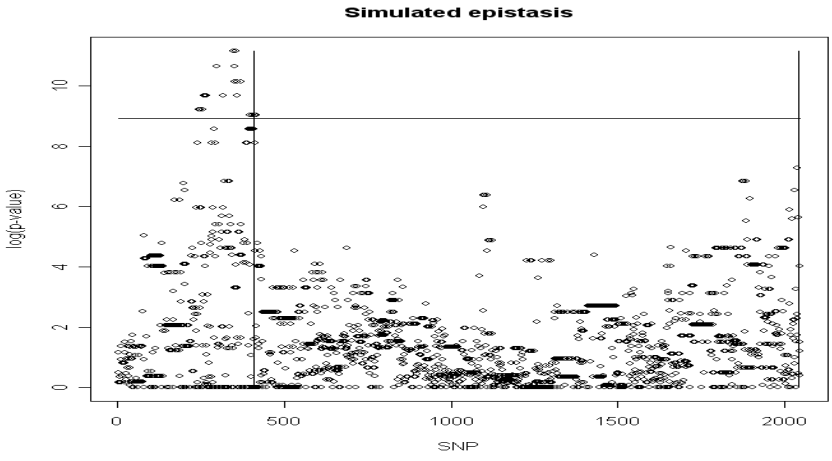
Fig. 2-Association plots of SNP markers for the simulated trait under scenario 1. Plots were obtained using 2 SNP haplotypes analyzed by a. SW/LR and b. ACA/LR. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations are significant at α=.05.



3 (a)

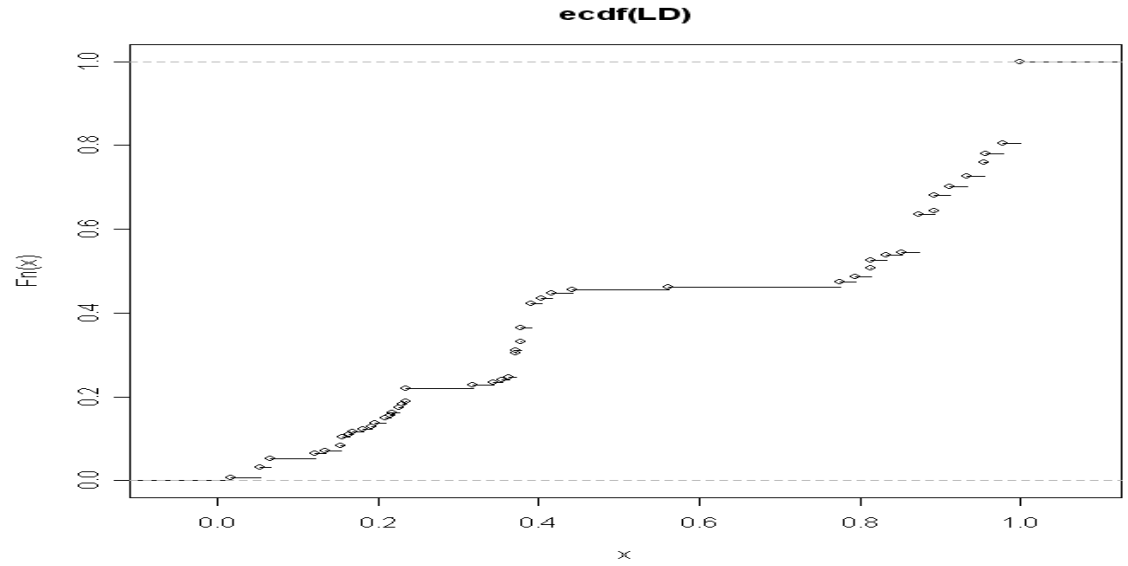


3 (b)

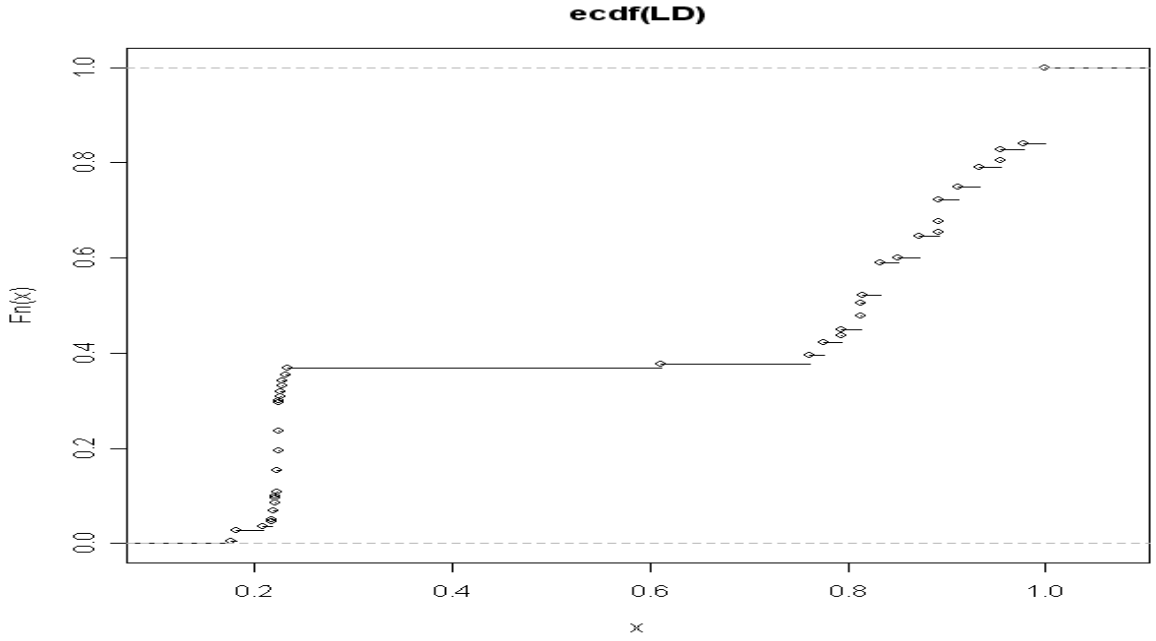


3 (c)

Fig. 3-Association plots of SNP markers for the simulated trait under scenario 2. Plots were obtained using 3 SNP haplotypes analyzed by a. SW/LR, b. ACA/LR, and c. RG. Vertical lines represent the position of the two causative mutations, and horizontal lines represent the threshold at which associations are significant at α=.05.



4 (a)



4 (b)

Fig.4-Plot of the cumulative distribution of SNP, identified as have significant associations when using a) ACA/G/D using 2 loci model (5.19%) b) RG/D , based on linkage disequilibrium with the causative mutations.

Contributors:

First and second authors contributed to the data simulation, data analysis. All three authors contribute to results interpretation and drafting. The second and third authors (co)directed all aspects of the study